ISSN: 2001-015X

Journal of Intelligence Studies in Business

Journal of Intelligence Studies in Business

Vol. 9, No. 2 2019

Included in this printed copy:

Editor-in-chief:
Klaus Solberg Søilen

Journal of Intelligence Studies in Business

Vol 9, No 2, 2019

The **Journal of Intelligence Studies in Business** (JISIB) is a double-blind peer reviewed, open access journal published by Halmstad University, Sweden. Its mission is to help facilitate and publish original research, conference proceedings and book reviews.

FOCUS AND SCOPE

The journal includes articles within areas such as Competitive Intelligence, Business Intelligence, Market Intelligence, Scientific and Technical Intelligence and Geo-economics. This means that the journal has a managerial as well as an applied technical side (Information Systems), as these are now well integrated in real life Business Intelligence solutions. By focusing on business applications, this journal does not compete directly with the journals that deal with library sciences or state and military intelligence studies. Topics within the selected study areas should show clear practical implications.

OPEN ACCESS

This journal provides immediate open access to its content on the principle that making research freely available to the public supports a greater global exchange of knowledge. There are no costs to authors for publication in the journal. This extends to processing charges (APCs) and submission charges.

COPYRIGHT NOTICE

Authors publishing in this journal agree to the following terms:

Authors retain copyright and grant the journal right of first publication with the work simultaneously licensed under a Creative Commons Attribution License that allows others to share the work with an acknowledgement of the work's authorship and initial publication in this journal. Authors are able to enter into separate, additional contractual arrangements for the non-exclusive distribution of the journal's published version of the work (e.g., post it to an institutional repository or publish it in a book), with an acknowledgement of its initial publication in this journal. Authors are permitted and encouraged to post their work online (e.g., in institutional repositories or on their website) prior to and during the submission process, as it can lead to productive exchanges, as well as earlier and greater citation of published work (See The Effect of Open Access.)

PUBLICATION ETHICS

The journal's ethic statement is based on COPE's Best Practice Guidelines for Journal Editors. It outlines the code of conduct for all authors, reviewers and editors involved in the production and publication of material in the journal. An unabridged version of the journal's ethics statement is available at https://ojs.hh.se/.

Publication decisions: The editor is responsible for deciding which of the articles submitted to the journal should be published. The editor may be guided by the policies of the journal's editorial board and constrained by such legal requirements as shall then be in force regarding libel, copyright infringement and plagiarism. The editor may confer with other editors or reviewers in making this decision. *Fair play*: An editor will evaluate manuscripts for their intellectual content without regard to race, gender, sexual orientation, religious belief, ethnic origin, citizenship, or political philosophy of the authors. *Confidentiality:* The editor and any editorial staff must not disclose any information about a submitted manuscript to anyone other than the corresponding author, reviewers, potential reviewers, other editorial advisers, and the publisher, as appropriate. *Disclosure and conflicts of interest*: Unpublished materials disclosed in a submitted manuscript must not be used in an editor's own research without the express written consent of the author.

Duties of Reviewers

Promptness: Any selected referee who feels unqualified to review the research reported in a manuscript, is aware of a personal conflict of interest, or knows that its prompt review will be impossible should notify the editor and excuse himself from the review process. *Confidentiality*: Any manuscripts received for review must be treated as confidential documents. *Standards of Objectivity*: Reviews should be conducted objectively. Referees should express their views clearly with supporting arguments. *Acknowledgement of Sources*: Reviewers should identify relevant published work that has not been cited by the authors. *Disclosure and Conflict of Interest*: Privileged information or ideas obtained through peer review must be kept confidential and not used for personal advantage.

Duties of Authors

Reporting standards: Authors of reports of original research should present an accurate account of the work performed as well as an objective discussion of its significance. Fraudulent or knowingly inaccurate statements constitute unethical behavior and are unacceptable. *Data Access and Retention*: Authors are asked to provide the raw data in connection with a paper for editorial review, and should be prepared to provide public access to such data (consistent with the ALPSP-STM Statement on Data and Databases). *Originality and Plagiarism*: The authors should ensure that they have written entirely original works, and if the authors have used the work and/or words of others that this has been appropriately cited or quoted. *Multiple, Redundant or Concurrent Publication*: An author should not publish manuscripts describing essentially the same research in more than one journal or primary publication. Submitting the same manuscript to more than one journal concurrently constitutes unethical publishing behaviour and is unacceptable. *Acknowledgement of Sources*: Proper acknowledgment of the work of others must always be given. *Authorship of the Paper*: Authorship should be limited to those who have made a significant contribution to the conception, design, execution, or interpretation of the reported study. The corresponding author should ensure that all appropriate co-authors and no inappropriate co-authors are included on the paper, and that all co-authors have seen and approved the final version of the paper and have agreed to its submission for publication. *Disclosure and Conflicts of Interest*: All authors should disclose in their manuscript any financial or other substantive conflict of interest that might be construed to influence the results or interpretation of their manuscript. All sources of financial support for the project should be disclosed. *Fundamental errors in published works*: When an author discovers a significant error or inaccuracy in his/her own published work, it is the author's obligation to promptly notify the journal editor or publisher and cooperate with the editor to retract or correct the paper.

ARCHIVING

This journal utilizes the LOCKSS system to create a distributed archiving system among participating libraries and permits those libraries to create permanent archives of the journal for purposes of preservation and restoration.

PUBLISHER

Halmstad University, Sweden
First published in 2011. ISSN: 2001-015X.
Owned by Adhou Communications AB

Journal of Intelligence Studies in Business

EDITORIAL TEAM

Editor-in-Chief
PROF KLAUS SOLBERG SØILEN (Sweden), Halmstad University

Founding Editors
PROF HENRI DOU (France), Groupe ESCEM
PROF PER JENSTER (China), NIMI

Honorary Editors
PROF JOHN E. PRESCOTT (USA), University of Pittsburgh
PROF BERNARD DOUSSET (France), Toulouse University

Regional Associated Editors
Africa
PROF ADELINE DU TOIT (South Africa), University of Johannesburg
America
PROF G SCOTT ERICKSON (USA), Ithaca College
Asia
PROF XINZHOU XIE (China), Beijing University
Europe
ASSOC PROF CHRISTOPHE BISSON (France), SKEMA Business School
Nordic
PROF SVEND HOLLENSEN (Denmark), University of South Denmark
PROF GORAN SVENSSON (Norway), Markedshøyskolen

EDITORIAL BOARD

PROF KARIM BAINA, École nationale supérieure d'informatique et d'analyse des systèmes, Morocco
DR EDUARDO FLORES BERMUDEZ, Bayer Schering Pharma AG, Germany
ASSOC PROF JONATHAN CALOF, Telfer School of Management, University of Ottawa, Canada
PROF BLAISE CRONIN, Indiana University, USA
DR SBNIR RANJAN DAS, University of Petroleum & Energy Studies, India
PROF HENRI JEAN-MARIE DOU, ATELIS Competitive Intelligence Work Room of the Groupe ESCEM, France
PROF BERNARD DOUSSET, Toulouse University, France
PROF ADELINE DU TOUT, University of Johannesburg, South Africa
PROF G SCOTT ERICKSON, Ithaca College, USA
PROF PERE ESCORSA, School of Industrial Engineering of Terrassa, Politechnical University of Catalonia, Spain
ASSOC PROF PER FRANKELIUS, Örebro University, Sweden
PROF BRIGITTE GAY, ESC-Toulouse, France
PROF MALEK GHENIMA, L'Université de la Manouba, Tunisia
PROF UWE HANNIG, Fachhochschule Ludwigshafen am Rhein, Germany
PROF MIKA HANNULA, Tampere University of Technology, Finland
PROF PER V JENSTER, Nordic International Management Institute, China
PROF SOPHIE LARIVET, Ecole Supérieure du Commerce Extérieur, Paris, France
PROF KINGO MCHOMBU, University of Namibia, Namibia
DR MICHAEL L NEUGARTEN, The College of Management, Rishon LeZion, Israel
PROF ALFREDO PASSOS, Fundação Getulio Vargas, Brazil
DR JOHN E PRESCOTT, University of Pittsburgh, USA
PROF SAHBI SIDOM, Université Nancy 2, France
PROF KAMEL SMAILI, Université Nancy 2, France
PROF KLAUS SOLBERG SØILEN, School of Business and Engineering, Halmstad University, Sweden
ASSOC PROF DIRK VRIENS, Radboud University, Netherlands
PROF XINZHOU XIE, Beijing Science and Technology Information Institute, China
DR MARK XU, University of Portsmouth, UK

MANAGERIAL BOARD

WAY CHEN, China Institute of Competitive Intelligence (CICI)
PHILIPPE A CLERC, Director of CI, Innovation & IT department,
Assembly of the French Chambers of Commerce and Industry, France
ALESSANDRO COMAI, Director of Miniera SL, Project leader in World-Class CI Function, Spain
PASCAL FRION, Director, Acrie Competitive Intelligence Network, France
HANS HEDIN, Hedin Intelligence & Strategy Consultancy, Sweden
RAÍNER E MICHAELI, Director Institute for Competitive Intelligence GmbH, Germany
MOURAD OUBRICH, President of CIEMS, Morocco

Copyright © 2019 JISIB, Halmstad University. All rights reserved.

Journal of Intelligence Studies in Business
Vol. 9, No 2 (2019) p. 4-5
Open Access: Freely available at: https://ojs.hh.se/

EDITOR'S NOTE **VOL 9, NO 2 (2019)**

Stevan Dedijer
1911-2004

A deeper look at the collective intelligence phenomenon

For the upcoming conference on Intelligence Studies at ICI 2020 in Bad Nauheim, Germany the focus of this issue of JISIB is on collective intelligence and foresight. The first two papers by Søilen and Almedia and Lesca deal with collective intelligence from an intelligence studies perspective. It may be said that the Internet itself is a gigantic collective intelligence effort, the largest in human history. Open source is a prerequisite for this system to work for everyone. The article by Černý et al. is on open source. All other contributions are on the connection between the Internet, software and intelligence.

This issue consists of seven articles to compensate for two articles that were taken out by editors in the last issue.

The first article by Søilen entitled "Making sense of the collective intelligence field: a review" is a historical review of the field of collective intelligence. The paper shows how collective intelligence is an interdisciplinary field and argues there is a flaw in the notion of "wisdom of crowds". Collective intelligence can be understood in terms of social systems theory and as such this approach has been fruitful for the social sciences, although so far not very popular. It also bares relevance for the study of business and economics.

The second article by Almeida and Lesca is entitled "Collective intelligence process to interpret weak signals and early warnings". Early warning and the detection of weak signals is a vital topic for any intelligence organization. Two aspects are discussed in the paper, the importance of new technology and collective sense making or interpretation

The third article by Shaikh and Singhal entitled "Study on the various intellectual property management strategies used and implemented by ICT firms for business intelligence" deals with intellectual property rights and patenting strategies. The authors identify a number of defensive and offensive IP strategies applied to ICT companies. The results have a bearing on patent acquisitions.

The fourth article by Lamrhari et al. is entitled "Web intelligence for understanding customer satisfaction: application of Latent Dirichlet Allocation (LDA) and the Kano model". Customer satisfaction today is mostly measured with data from the internet, using different business intelligence techniques. The Kano model is still valuable[i,ii], but the way we gather information to assess the different levels in the model has changed. The authors use Latent Dirichlet Allocation to analyze the voice of customer (VOC) in online reviews. They suggest that BI techniques and a fuzzy-Kano model can enable companies to better understand their customers' online reviews.

The fifth article by Nahili et al. is entitled "A new corpus-based convolutional neutral network for big data text analysis". Companies need efficient ways to analyze everything that is said about them on the internet (reviews, comments). The paper suggests a convolutional neural network (CNN) as it has been successfully used for text classification. IMDB movie reviews and Reuters datasets were used for the experiment.

The sixth article by Černý et al. is entitled "Using open data and google search data for competitive intelligence analysis". Taking the Czech antidepressant market as an example, the authors show how competitive intelligence can be obtained using Google Search data, Google Trend and other OSINT sources.

The seventh article by Dadkhah et al. is entitled "The potential of business intelligence tools for expert findings". The paper suggests a way for researchers to find experts using business intelligence tools. The same method may also be used by any business or person looking for experts on a specific topic.

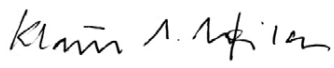
As always, we would above all like to thank the authors for their contributions to this issue of JISIB. Thanks to Dr. Allison Perrigo for reviewing English grammar and helping with layout design for all articles and to the Swedish Research Council for continuous financial support.

Copyright © 2019 JISIB, Halmstad University. All rights reserved.

We hope to see you all at the ICI 2020 on the 16-17 March, 2020. The deadline for the two-page abstract submission is March 1st, 2020.

On behalf of the Editorial Board,

Sincerely Yours,



Prof. Dr. Klaus Solberg Søilen
Halmstad University, Sweden
Editor-in-chief

Copyright © 2019 JISIB, Halmstad University. All rights reserved.

[i] Tontini, G., Solberg Søilen, K., Silveira, A. (2013). How interactions of service attributes affect customer satisfaction: A study of the Kano model's attributes. Total Quality Management & Business Excellence, 24(11-12), 1253-1271

[ii] Tontini, G., Söilen, K. S., & Zanchett, R. (2017). Nonlinear antecedents of customer satisfaction and loyalty in third-party logistics services (3PL). Asia Pacific Journal of Marketing and Logistics, 29(5), 1116-1135.

Journal of Intelligence Studies in Business
Vol. 9, No. 2 (2019) pp. 6-18
Open Access: Freely available at: https://ojs.hh.se/

Making sense of the collective intelligence field: A review

Klaus Solberg Søilen[a]

[a]*Department of Engineering, Natural Sciences and Economics, Faculty of Marketing, Halmstad University, Halmstad, Sweden*

**Corresponding author: klasol@hh.se*

Received 3 June 2019 Accepted 15 September 2019

"The world is bitterly, savagely competitive and intensely, vigorously cooperative, by way of alliances and partnerships, thus rapidly changing individuals and social systems alike."

"We are pulled toward a single social system on earth."

Dedijer, 1999, p. 72

ABSTRACT The problem we want to solve is to find out what is new in the collective intelligence literature and how it is to be understood alongside other social science disciplines. The reason it is important is that collective intelligence and problems of collaboration seem familiar in the social sciences but do not necessarily fit into any of the established disciplines. Also, collective intelligence is often associated with the notion of wisdom of crowds, which demands scrutiny. We found that the collective intelligence field is valuable, truly interdisciplinary, and part of a paradigm shift in the social sciences. However, the content is not new, as suggested by the comparison with social intelligence, which is often uncritical and lacking in the data it shows and that the notion of the wisdom of crowds is misleading (RQ1). The study of social systems is still highly relevant for social scientists and scholars of collective intelligence as an alternative methodology to more traditional social science paradigms as found, for example, in the study of business or management (RQ2).

KEYWORDS Collective intelligence, social intelligence, social systems, wisdom of crowds

1. INTRODUCTION

The popularity of the collective intelligence research area has increased significantly. The Web of Science lists 552 article with the term in the title, the first of which was written in 1989. The last 500 articles were written since 2005. Research groups at the most prestigious universities receive grants to establish separate research centers and the ideas have received significant interest from the general public as well as politicians. At the same time the phenomenon seems old and familiar in the scientific literature. Moreover, the field seems to be highly interdisciplinary and does not seem to fit into any of the established business, management or social sciences disciplines. So, what is new and valuable in this field of how we learn and make decisions together? (RQ1). In addition, how are we to understand where collective intelligence fits in a larger social science context? (RQ2) The research gap suggests that there is no critical review article

that examines the phenomenon of collective intelligence from a historical context where the aim is to understand what this body of literature is about.

2. METHOD

This article attempts to answer the research question through the historical method, comparing what has been written in the past about learning together to the spread of collective intelligence during our own time. Moreover, the attempt is to compare the collective intelligence literature to that of social intelligence. Social intelligence was present in the 1970s, at the start of what became intelligence studies in business. The sources are scientific articles, books, internet articles and videos. I have attempted to follow a theme, inevitably missing much relevant information as the phenomenon is so wide and spread over synonyms containing the words intelligence, collaborative, collective, crowd, group, knowledge, open source, smart, social, and connectivity, just to mention some of the most relevant. The methodological problem here is first one of what articles to select and why. I have chosen to read the most cited articles first, the most popular non-scientific sources and what can be deemed significant scientific contributions over time, including books. This limited the sources down to less than fifty relevant publications, where about half are listed as references here. In terms of scientific articles, there were about thirty that had twenty or more citations in the Web of Science. All of them have been included here. I have not cited sources I have not read in their entirety. Only a few have been discarded, as they were too technical.

There are numerous limitations in this study. Leading articles and leading scholars are reduced to citations on Web of Science and Google Scholar, which does not give the full picture. Further, it would be interesting to go deeper into each of the disciplines mentioned in the articles, both when it comes to definitions, but more important to their actual meaning and content to detect similarities and differences, but also to investigate the theories and experience they build on. Part of this is due to the limited number of pages allowed in the article by the journal.

3. LITERATURE REVIEW

In 1886 Francis Galton, a cousin of Charles Darwin, wrote an article called "Regression towards mediocrity in hereditary stature" which showed that there was a regression towards the mean with larger numbers. This was statistically proven by for example having a large number of people guess the weight of an ox at a fair. As the number of responses increased, the average guess ended up reflecting the actual weight, showing a simple linear regression of data points. The technique was useful for simple questions demanding numerical answers, but Galton thought, as the title suggests, that the logic would lead to "mediocracy" when applied to other problems. This critique was considered common sense at the time, supported by scientists, humanists and men of letter alike (from Henry David Thoreau to Friedrich Nietzsche).

However, despite the critique, the idea was useful in statistics and received renewed attention with the rise of computer science and in particular big data and now with artificial intelligence and digital marketing, for example when counting averages such as webpages visited or number of clicks on a webpage. This tells us about peoples' behavior online. Web 2.0 caught on during the first decade of the new millennium, the idea of creating content through interaction and collaboration using social media. In rapid succession, Facebook was founded in 2004, YouTube the year after and Twitter the year after that. When competitors arrived, they were simply bought up, guaranteeing near-monopolies for the new data giants. Due to the large amount of data collected, these companies are now able to predict our behavior more accurately as they collect more data on and from us.

Researchers saw this development coming, thus in 2002 Howard Rheingold argued that the most successful services in the future would not be hardware devices or software programs, but social practices online. In 2004 a young American journalist James Surowiecki wrote a book with the provocative title "The Wisdom of Crowds", based on Galton's idea. However, Surowiecki takes the idea further delving into economics, rejecting Adam Smith and other economists for their focus on specialization. Instead, he argues for decentralization:

> "Decentralization's great strength is that it encourages independence and specialization on the one hand while still allowing people to coordinate their activities and solve difficult problems on the other". (P. 71).

In other words, valuable information may not come through when only a few are in the know. He gives the example of the CIA. The original idea of having a centralize intelligence agency as defined by Bill Donovan was later abandoned as the agency grew and more departments were established. These departments did not succeed in cooperating and sharing information, a consequence of which was the attacks on September 11th, 2001. The problem was timely and the book became a bestseller.

> "The Congressional Joint Inquiry into the attacks found that the U.S. intelligence community had 'failed to capitalize on both the individual and collective significance of available information that appears relevant to the events of September 11.' Intelligence agencies 'missed opportunities to disrupt the September 11th plot,' and allowed information to pass by unnoticed that, if appreciated, would have 'greatly enhanced its chances of uncovering and preventing' the attacks. It was, in other words, Pearl Harbor all over again." (Surowiecki, 2004; P. 68)

Surowiecki draws a parallel to Galton's contributions, but a critic may argue that the information workers at the CIA are not your average visitor to the fair guessing the weight of an ox. The author is mixing experts and professionals with average people. Quiz games are a good counter example; you only stand a chance of winning if you can manage to gather knowledgeable people on your team. If you make up the team with those who just happen to walk into the pub that evening your team will have a small chance of winning. Lanier (2006) notes that the collective is more likely to be smart only when:

1. It is not defining its own questions,
2. The goodness of an answer can be evaluated by a simple result (such as a single numeric value), and
3. The information system, which informs the collective, is filtered by a quality control mechanism that relies on individuals to a high degree.

Lanier argues that only under those circumstances can a collective be smarter than one person. If any of these conditions are broken, the collective becomes unreliable or worse." (Wikipedia).

Another critical point is made by Tammet (2009), who argue that in systems of pooling knowledge, like Wikipedia, experts can be overruled by less knowledgeable persons. Thus it is important to build software that immediately alerts the experts when changes to the entry are made and allow discussion on the issues, saving these for other users to partake in to judge who is right. To build this system as a Galton-average-towards-the-mean would not work. In other words, Wikipedia works well because it pools smart people, despite the disturbance of less smart individuals because there are special mechanisms built into the system to deal with their erroneous entries.

Maybe the best counter argument was a game of chess held in 1999 called "Kasparov versus the world", where the chess player played against over 50 000 people from more than 75 countries deciding moves by plurality vote. An expert system was put in place whereby four highly rated players (FIDE ranking) suggested moves first. These suggestions were mostly followed by 'the world'. Kasparov won despite the experts, but he admitted it had been a tough match. If Kasparov had played against an average move we can assume that he would have won easily. Instead it must be suggested that the wisdom in the crowd is a romantic idea that fits well with the reigning democratic political ideology in the Western world and the equally dangerous belief that advancements in computer science will solve collective problems. It will certainly solve some, but new dangers will arise, as we saw with the invention of nuclear energy.

Surowieci is right when he says that "The idea of collective intelligence helps explain why, when you go to the convenience store in search of milk at two in the morning, there is a carton of milk waiting there for you, and it even tells us something important about why people pay their taxes and help coach Little League." (P. XIV), but not of the reasons he describes. There is milk in the store because the store managers knows how many customers buy milk on a specific weekday. The more of his business he can digitize the better information he will have on customer's' behavior. His other example is that many people pay their taxes because they know that it benefits all in society including themselves, especially as they get older. Of course, most pay taxes because they

have to and do what they can to avoid paying them. So, these are not good examples of what the author wants to convey.

Looking at research during the past decade: Among the more cited research articles in the field are Woolley et al. (2010), presenting a short empirical experiment, where they found that social sensitivity and proportion of females explains why some groups work better together. In experiments like these, it's difficult to know what are the causes and effects, and it may be that IQ or other factors are better explanatory variables. Woolley et al. publish another article in 2015 with the same test, but it's difficult even to assess this one as it's short and does not describe the method, analyses or show data. Engel et al. (2014) argue that the same findings are just as true in online environments. The authors define collective intelligence as "the ability of a group to perform a wide variety of tasks" or "the general ability of a particular group to perform well across a wide range of different tasks". This is different from other definitions, for example as defined in Wikipedia: "the intelligence that emerges from collaboration, collective efforts and competition among individuals".

Furthermore, there is an understanding in these articles that collective intelligence implies that the sum of the efforts from all individuals in the group are greater than the sum of each individual's contribution, so that 2+2=5, as it were. This is an attractive idea, but there are no good empirical experiments that confirm this assumption. It may be true in some cases, as when members of a quiz team only know parts of an answer each but become convinced when they pool their arguments together, making a strong case for a specific idea, but then again we are dealing with experts not with the average person.

There is one mathematical paper that addresses this problem. Nguyen (2008) shows how the intelligence of a collective can be larger than the intelligence of its members through mathematical modelling. "These examples show that the relationship between the intelligence of a collective and the intelligences of its members is not linear" (P. 543). "Thus, with some restrictions, one can claim that the hypothesis A collective is more intelligent than one single member is true." (P. 561). However, the paper builds on the implicit assumption that every member knows the same and for example is not wrong on a specific issue, which can cause confusion in a group. Knowing this the assumptions can hardly be said to be realistic when dealing with crowds. It is the same *certeris parabus* we find behind most of what has been written about economics since the Second World War, we assume that all rational individuals can weigh alternatives and draw the right conclusions based on them. Individual and cultural differences (reality) tend to destroy most of these social science models. We can also say, it's the weakness of linear logic.

When looking at videos on collective intelligence, bees and ants are often used as analogies to show what can be achieved in the social sciences. There is both substantial and interesting research on the behavior of bees and ants performed by natural scientists. The first time the term 'collective intelligence' appears in research is in a study of ants (Franks, 1989). "The sharing and collective processing of information by certain insect societies is one of the reasons that they warrant the superlative epithet 'super-organisms' (Franks 1989, p. 138)." But the comparisons between species, even different kinds of bees, are more complicated, as Franks et al. remind us of in an article from 2002:

> "Nevertheless, both species do make use of forms of opinion polling. For example, scout bees that have formerly danced for a certain site cease such advertising and monitor the dances of others at random. That is, they act without prejudice. They neither favour nor disdain dancers that advocate the site they had formerly advertised or the alternatives. Thus, in general the bees are less well informed than they would be if they systematically monitored dances for alternative sites rather than spending their time reprocessing information they already have." (P. 1583)

More to the point, people are not bees or ants and no one would like to be one, I believe, or to live according to their motives. This comparison is what is thought of as a mechanical worldview in the business literature. At the end it brings associations to fascism, hardly an attractive metaphor. Instead we as human beings enjoy our irrationalities, our cumbersome ways even our flaws. It is part of what makes us human. This is no denying that human are animals, but our behavior seem to be substantially different from those of ants and bees in general making the parallels of limited value.

The most cited article on collective intelligence and honeybees by Rajasekhar et al. (2017), argue that the algorithms developed over the past twenty years to understand their behavior are not well adapted to real life problems. The authors refer to an article by Sörensen (2015), who express his concern on the current trends in metaheuristic research (i.e. higher-level procedure or heuristic designed to find, generate, or select a method for solving problem) in the following way "… it seems that no idea is too far-fetched to serve as inspiration to launch yet another meta-heuristic. …we will argue that this line of research is threatening to lead the area of metaheuristics away from scientific rigor". "The ideas should be presented in a metaphor-free language and more directly" (in Rajasekhar, 2017; P. 45).

In everyday business life a good collective intelligence system is developed as some sort of a business intelligence software. Thus valuable contributions to the field of collective intelligence will continue to come from software development. This is a continuation of web 2.0, a comparison which has its own problems:

> "The most hyped examples of collective intelligence applications have been labeled as "Web 2.0" applications. Web 2.0 is an amorphous term used to define a computing paradigm that uses the Web as the application platform and facilitates collaboration and information sharing between users" (Gregg, 2010; P. 134). "The shift to a collective intelligence paradigm requires software developers to have different ways of thinking about how their how software might be used and what features would enable better visualization and use of information among groups of people. The new breed of collective intelligence applications needs to center around user defined data that can be reused to support decision making, team building, or to improve understanding of the world around us." (P. 134).

Collective intelligence in this sense and for this group of researchers means developing new and better business intelligence software for collaboration.

Lykourentzou et al. (2010) sees collective intelligence as a continuation of a wiki. The authors present what they call a CorpWiki, "a self-regulating wiki system for effective acquisition of high-quality knowledge content" (P. 18). "Inserted articles undergo a quality assessment control by a large number of corporate peer employees. ". This is close to the description of a software the author of this paper developed in 2004 called Subsoft, which never made it passed a beta version but was tested in local government organizations, not that it was unique.

The core research question of the Center for Collective Intelligence at MIT is "How can people and computers be connected so that – collectively – they act more intelligently than any individuals, groups, or computers have ever done before?" (Leimeister, 2010). This understanding is not that different from how software developers work. Software is not developed in a vacuum but with the users' needs in mind, users who become ever more collaborative. The software simply reflects this reality with continual technological discoveries, giving rise to new product developments.

Just as with the effort to advocate for open source in software development, there are efforts to influence how collective intelligence systems are made, so as to make them more beneficial for all. We are now in the domain of political science and law. Schum et al. (2012) argue that the software should not be restricted to "government, scientific or corporate elites, but be opened up for societal engagement and critique" (P. 110). Basically, what is suggested is not that different from Wikipedia, but with some policy improvements on criteria: There should be:

> "transparency of data sources, algorithms, and platform use – control of users over their personal data – privacy-respecting data mining – self-regulation, self-healing – reliability and resilience – promotion of constructive social norms and responsible use – crowd-based monitoring of platform use, involving non-profit organizations – tools to alert problems and conflicts, and to help solving them – incentives to share profits generated from data and algorithms provided by users – mechanisms for managing unethical use." (P. 112-113).

Thus, we may already make our first conclusion: that the body of literature published under the collective intelligence umbrella is truly interdisciplinary (Conclusion # 1).

Wolf et al. (2015), tests the ideas of collective intelligence to increase decision accuracy on medical decision-making. The authors found that "all CI-rules systematically outperform even the best-performing individual radiologist in the respective group", and that "the findings demonstrate that CI can be employed to improve mammography screening". (P. 1). Again, in this case it's experts - "multiple radiologists" - who give their input. These experiments do not confirm Galton's regression towards the mean but the fact that many experts perform better than one, which is common sense, but also costly and thus less practical in real life. A more promising solution to this problem seems to be artificial intelligence, using computers instead of humans, but that is for another paper on a different topic.

A second conclusion is that we are confronted with the phenomenon we may call wisdom of the knowledgeable more than wisdom of the crowds (poking fun at Surowiecki, who in turn pokes fun of Charles Mackay's article about the "Madness of the crowds". See Mackay, 1841). The logic of crowds works for problems of how much an ox weighs or what the consumption of milk may be tomorrow, but not that well on problems of how to win a quiz tournament, or, closer to home, what goes on in a company or how to understand an industry. If we ask what the capital of Senegal is we may get the correct answer among thousands of answers, but how are we to know which one to choose if we are not allowed to check with someone who is smarter, more knowledgeable than the rest (Conclusion # 2).

Wisdom of the knowledgeable is common sense thus a less interesting conclusion. It is not the kind of title to sell books. What we can say is that the observation is reasonable and confirms what we have known for a very long time. There is another problematic aspect of the term 'wisdom of the knowledgeable' and that is the question of whether the knowledgeable are truly wise. The wise make decisions based on what is best from the wider perspective, in the long run. Being knowledgeable by no means guarantees that we are wise. Our modern society is becoming ever more short-term focused (financial markets, profits, product life cycles, etc.), increasing the gap between wisdom and knowledge. Another way of saying this is that neither the crowd nor the knowledgeable seem very wise. (Conclusion # 3).

The next question to consider is whether the literature reviewed on collective intelligence literature is new. The phenomenon studied is part of the topics studied under what we call the information age, preceding the industrial revolution. Alvin Toffler was one of the pioneers in the digital revolution of the 1970s and 1980s (Toffler, 1980).

Stevan Dedijer, a contemporary of Toffler, wrote more specifically on intelligence and developed what we call social intelligence. His predecessor at the University of Lund, Wilhelm Agrell, explains in a foreword:

> "Central to his work, his reading and vast correspondence was a concept of what he called social intelligence: the ability of individuals and organizations to orientate in an increasingly complex information environment… Stevan foresaw the coming of an age where individuals and organizations alike would become dependent on this ability to collect, process and use information curiosity and insights information and the immense challenge of a coming information explosion" (p. 7) (Dedijer, 1999)

Dedijer was well aware of the contributions that had preceded his own work. "If we look back before Web of Science and other databases collected that many articles the first insights of 'organized intelligence,' 'social intelligence,' and of a 'planetary intelligence sphere' emerged in the 1920s." (Dedijer, 1999, p. 69). "Like Mendel's article in 1903 on genetics, they were totally ignored for decades. Walter Lippman advocated in his 'Public Opinion' (1922) the use of 'organized intelligence' in all fields of government. The philosopher John Dewey in the l930s saw 'organized and social intelligence' as the only tool humanity could use to avoid the Scylla of totalitarianism and the Charybdis of laissez-faire market capitalism." (p. 69). Dedijer observed the changes that intelligence was brining during his own time: "The basic intelligence goal for individual countries is changing from intelligence for national security to intelligence for national growth and development." (p. 67). As such, he also foresaw the change from geopolitics to geoeconomics that Luttwak wrote about (Luttwak, 1990) and he foresaw that mass communication would lead to "individualization of intelligence", with users becoming more isolated, self-centered, and egotistic. The crowd would get louder, more

daring in its attack. We see this on social media today with the phenomenon of trolls, spilling over to populism and the weakening (not strengthening) of the democratic process (as is implicit in the "wisdom of crowds").

Dedijer, who fought in the US military as a paratrooper during the Second World War, worked on question of intelligence with the CIA and W. Colby, its director. The two friends shared information about how they saw the world changing and how the intelligence services should adapt. One of the developments Colby did not anticipate was the importance of collaboration:

> "The second dimension I added to Colby's intelligence 'elephant' was the emergence of development sciences related to the individual, various social systems, and humanity in general. All are engaged in 'bridge building' among biological, individual, social, technological, and global intelligence and social systems." (P. 70). "'Bridge building' [– what we call interdisciplinary today -] is the name for current attempts at a holistic approach to all kinds of problems in every discipline or field. One of the best formulations of the bridge-building method is found in mathematics. S. Singh in Fermat's Enigma: The Epic Quest to Solve the World's Greatest Mathematical Problem (l998) tells how A. Wiles proved in l995 a conjecture that confounded the greatest mathematicians for 358 years: 'Mathematics consists of islands of knowledge...each one with its own unique language, incomprehensible to the inhabitants of other islands... Mathematicians love to build bridges. The value of mathematical bridges is enormous. They enable communities of mathematicians who have been living on separate islands to exchange ideas and explore each other's creations.' Such bridge-building techniques are used in physics, as shown by Nobel Laureate S. Weinberg in the development of individuals as well as social systems, including studies of the state of humanity." (P. 70).

Interdisciplinarity of social systems was developed simultaneously, it seems, by a number of people, among whom the more influential included the German philosopher Niklas Luhmann (1968 and 1984), Kenneth Boulding (1956) and Ackoff (1971) in the US.

Dedijer believed that the intelligence discipline was going to be valuable for the social sciences, but he also saw the difficulties the discipline was facing due to its unfortunate parallel and association to spying.

> "Because of isolation and confusion among intelligence disciplines and the myth that intelligence is above all espionage, billions of individuals, organizations, and governments today use information technology yet fail to perceive the innumerable signals which tell of a new intelligence revolution in the evolution of humanity." (Dedijer, 1999, P. 71).

This is a problem that the collective intelligence literature is also confronted with, by default so to speak, as will any new discipline that uses the term intelligence more in the sense of 'information' than 'brains'.

In conclusion, we have shown that collaboration and sharing of information was at the heart of Dedijer's idea of social intelligence. We argue that both collective intelligence and social intelligence is part of the same paradigm shift, like two waves of the same current. Just like AI has come and gone with new enthusiasm and interest the past decades, so the 'information turn' is visited and revisited with certain intervals and different approaches. We shall understand all of these developments as part of an ongoing intelligence paradigm. This is our forth conclusion (Conclusion 4).

The term 'intelligence paradigm' can be related to systems thinking, as will be discussed further in the analysis below. The term is also used by Lahneman (2010) related to international politics and security, and by Zadeh, (2008), related to machine learning, but we shall keep these two tracks out.

4. ANALYSIS OF THE INTELLIGENCE PARADIGM AS SYSTEMS THINKING

Kuhn (1962) defined paradigm rather broadly as a development that "designates what the members of a certain scientific community have in common, that is to say, the whole of techniques, patents and values shared by the members of the community". According to this broad definition there could be hundreds if not thousands of paradigms just in the study of economics and management alone.

Ackoff (1971) writes about the paradigm shift required for the study of management to redirect to systems thinking, referred to as complex systems and complexity theory. The basic idea is that organizations stop thinking of themselves divided into sections such as marketing, HRM, and strategy, but instead as elements that form relationships. It's the connectivity of the parts that is valuable, not the parts themselves. Ackoff's favorite example is the car. All the parts by themselves are useless, even added together as a sum they give nothing. It's the right connectivity of the parts that give an automobile that is actually useful and can take us from point a to b. The principles governing how we run business organizations should not primarily be existing departments but the exchange of information, or intelligence. In other words, the private organization is best run as an intelligence organization, much like state intelligence institutions. Many successful private organizations today do just that, like the largest wealth management fund in the world, Blackrock. Its offices and data facilities remind one more of the NSA than a classic bank. Most major companies today look much the same, including Google, Facebook and Amazon. The success they achieve is primarily determined by the value of the information they gather and analyze. Whether we as employees work in marketing or HR we are spending more and more time learning about new computer systems, electronic gadgets and related services. Without these skills we are worth little on the labor market.

One problem is that universities and learning institutions often assume that students already know this. The individual disciplines (economics, marketing, HR) are not taking into consideration how these new technologies are changing professions. One example is marketing. Students do not know digital marketing when they come to university. Actually, that is what they come to learn. If the teacher assumes that these are skills that the students already know and that it's enough to teach a broad set of general theories, then the education fails.

In reality, we have all become information workers during the past generation. The major difference today seem to be that some build the systems (engineers) and others use them (engineers and everyone else). Knowledge and skills have never been as important as now. Even to work in a factory you need more than a high school diploma. Never before in the history of mankind have companies been better at locating knowledgeable people and bringing them together, no matter where they are on the planet. This development matches poorly with the notion of wisdom of the crowd. Companies are not hiring just anybody, but are getting better at finding those few who possess supervisor knowledge and experience. There is nothing appealing about the crowd except that all customers of the same product are worth just as much in terms of money (economic reasoning) and that one human life is not worth more than another (our shared human value).

Instead, the notion of wisdom of the crowd is appealing for political reasons, because it supports the notion that all citizens have a say and can control their own future through democratic elections, which is the basis of Western societies. Western governments support these ideas because it strengthens the status quo. In the same way, wisdom of the knowledgeable, besides being obvious as a term, thus dull, sounds elitist. The notion of wisdom of the knowledgeable brings up a painful contradiction in Western civilization. It indicates a difference between democratic and meritocratic values, which is as old as Western democracy and has been actively debated in Europe since the early 1960s (Young, 1959). To understand the popularity of collective intelligence it's impossible to ignore these political aspects. Politics may be the single most decisive factor for shift in scientific paradigms, not for having the ideas, but getting them implemented.

For this reason, it shall be suggested that the intelligence paradigm shift is probably not going to come from the Western world, but from Asia. The Asian way of conducting business and working is already in many ways similar to an intelligence approach. Chinese companies thrive by learning from the West, by travelling to foreign countries and copying our products. The whole Belt and Road Initiative (BRI) is a gigantic collective and collaborative effort in the spirit of the Competitive Advantage of Nations, an idea we used to master but have forgotten. As a result, it's not we who know more about Asia than they about us but the exact opposite: Our students know next to nothing about them, while their students know much about us, and are keen learners.

Asian companies are not limited by compartmentalized knowledge. Instead, they look for useful knowledge where they can find it (what works) and are in many ways better at solving problems. The popular notion is that

this is what we are good at, because we are more used to, or allowed to, question things. It was what the Western world did well after the enlightenment. Since then we have become less curious about the world, less eager to change it and instead more concerned with our own immediate private needs. A tragic example is that our social media applications have made us more isolated, not more collaborative. These services have made us less knowledgeable about the world, not more.

Dedijer understood this danger well as a leading nuclear physicist: "Information Technology is only a tool. Always ask how effective and efficient it is in terms of improving your capability to identify and solve problems by acquiring and using the information it can help to provide. The IT model of the future will more and more be "a thing that thinks"', as we call artificial intelligence. AI is further away from being a reality than what we are led to think, where the delay in self-driving vehicles is just a reminder.

This may be the real difference from Dedijer's social intelligence to Surowiecki's collective intelligence, that now we are discovering machines that can "think" (artificial intelligence): more effective, more interactive, and faster IT systems that makes it easier to learn together. It is the study of how this is happening that lies at the core of collective intelligence. It is a world of new opportunities brought forward primarily by computer scientists and neuroscientists, but where social scientist will play an important part in evaluating applications and consequences. For this the literature will need be more critical. (Conclusion # 5). As the example of the Facebook–Cambridge Analytica data scandal has confirmed, social scientists should not be a gospel choir in the church of progress.

The age of information is changing everyone's lives. Writing this research article is collective intelligence made possible by information technology, especially large databases (Web of Science) and fast internet connections (from home, or on the train on my way to work). Instead of meeting colleagues and exchanging information on a topic, we write articles and share them. I try to locate those who know more than me and learn from them. That is an active process of collective intelligence.

The idea of collective intelligence is as old as mankind, as man quickly discovered that he had to cooperate and pool ideas if he wanted to trap and kill larger animals like the mammoth. The notion has been a frequent topic in literature throughout time to the point where it is difficult to say who has contributed the most to it.

The literature on collective intelligence is a good example of non-collaboration. Ever greater specialization in the social sciences draws groups of scientists and researchers further apart even when they study the same phenomenon. The reason this happens is because the databases we use do not contain older articles (basically just the last fifty years), there are almost no articles in other languages than English (even though much progress was communicated in German and French), and researchers come up with new buzz words to establish their own careers and distinguish themselves from others, for personal and economic reasons. If the social science project was truly critical, this reinvention of the wheel should not be possible. In the German scholarly tradition, one is always confronted with the question of meaning. "What does that mean?", with the clear goal of understanding a phenomenon. Due to a systematic lack of such questions and aims, in the social sciences we now have dozens of groups, or tribes, studying the same phenomenon: artificial intelligence, collective intelligence, information sciences, and intelligence studies. The difference is the size of these groups, what networks they belong to and their financing. There are of course also differences in relevance and output of research.

The larger question is if questions of collaboration will continue to be studied by multiple disciplines with little contact between them, or if the modern social science project will merge into something else. Stevan Dedijer suggested social systems theory, going back to Bertalanffy (1968), and he explains:

> "The world is bitterly, savagely competitive and intensely, vigorously cooperative, by way of alliances and partnerships, thus rapidly changing individuals and social systems alike… We are pulled toward a single social system on earth." (Dedijer, 1999, P. 72).

Others, have elaborated the idea further. Mainzer concludes that [we must]

> "learn to consider humans as complex nonlinear entities of mind and body… the

> theory of complex systems explain what we can know and what we cannot know about nonlinear dynamics in nature and society… we need to 'improve our knowledge of complexity and evolution'… mono-causality often leads to dogmatism, intolerance and fanaticism" (Mainzer, P. 294-5)

The same basic idea from the social systems literature in the social sciences is found in the complex systems literature in the natural sciences and in information sciences: behavior cannot easily be studied with small (for example, student group surveys), narrow (a few isolated variables) empirical projects with data of short duration (behavior changes in time and depending on circumstances). It requires the complexity of a multifaceted social structure. Any modelling that tries to reduce reality to a correlation analysis performed on a few variables is of limited use. But, do leading scholars interested in collective intelligence interest themselves for systems thinking and complex systems today? Yes, they do.

4.1 Analysis of Research Areas

To find out we analyzed the top-ranking scholars on collective intelligence according to Google Scholar, seeking out those with 1500 or more references. These are listed anonymously in the table according to their respective ranking. A total of five keywords or research topics are possible on Google Scholar, where it is common (but not certain) to list them according to the main interest of the researcher.

Five of the leading scholars are focusing on complex systems. That is more than for any other research area. Two of the four leading mention complex systems as a specialty.

Table 1 ***Keywords associated with the leading scholars on collective intelligence, according to Google Scholar****. The columns show areas of study, ranked according to each person's interest. The individual scholars are listed anonymously by ranking.*

Rank	Primary	Secondary	Tertiary	Quaternary	Quinary
1	Artificial Intelligence	Ontology	Collective Intelligence	Virtual Assistants	Intelligent Interfaces
2	Intelligence Augmentation	Collective Intelligence	Open Science	Quantum Information	Quantum Computing
3	Collective Behaviour	Collective Behavior	Swarm Intelligence	Collective Intelligence	Complex Systems
4	Machine Learning	Complex Systems	Data Mining	Information Retrieval	Collective Intelligence
5	Democracy Innovation	Innovation	Technology	Collective Intelligence	
6	Computational Creativity	Collective Intelligence			
7	Self-Organization	Collective Intelligence	Cybernetics	Complex Adaptive Systems	Distributed Cognition
8	Learning Analytics	Argument Mapping	Collective Intelligence	Human-Computer Interaction	
9	Knowledge Engineering	Collective Intelligence			
10	Collective Intelligence	Artificial Intelligence	Multi-Agent Systems	Sustainability	
11	Information Systems Design	Design	Visualization	Crowd Work	Collective Intelligence
12	Artificial Intelligence	Collective Intelligence	Cultural Algorithms	Evolutionary Computation	
13	Biological Physics	Statistical Physics	Slime Molds	Networks	Collective Intelligence
14	Social Decision Making	Collective Intelligence	Empathy	Justice	
15	Artificial Intelligence	Collective Intelligence	Human-Computer Interaction		
16	Collective Behaviors	Crowds	Computational Social Science	Complex Systems	Collective Intelligence
17	Swarm Intelligence	Collective Behavior	Collective Intelligence	Social Behavior	Slime Molds
18	Neuroscience	Psychology	Education	Collective Intelligence	Aging
19	Population Dynamics	Social Systems	Collective Intelligence		
20	Global Futures Research	Foresight	Futures Research Methodology	Global Challenges	Collective Intelligence
21	Business Analytics	Data Science	Crowdsourcing	Collective Intelligence	
22	Information Systems	Network Science	Computational Social Science	Crisis Informatics	Collective Intelligence
23	Network Science	Collective Intelligence	Crowdsourcing		
24	Systems Biology	Synthetic Biology	Statistical Inference	Self-Organization	Collective Intelligence
25	Democratic Theory	Constitutional Theory	Political Epistemology	Philosophy Of Science	Collective Intelligence
26	Computational Intelligence	Collective Intelligence	Natural Language Processing	Machine Learning	
27	Digital Innovation	Open Innovation	Collective Intelligence	Complexity	Computational Social Science

From the data we also learn that collective intelligence only appears once in the first position. On average, it is in the fourth place, which means that it is not a priority even for those who focus on this area. Artificial intelligence is the most reoccurring specialization, occurring three times in first place. Collective behavior is mentioned two times. The large majority of co-subjects are technical, related to information sciences at large, with few contributions from the social sciences. The variety of technical specialization is very large too. Topics related to crowds occur four times in total, innovation four times. We conclude that the direction of complex systems as a way to study the social sciences, and problems of collective intelligence in particular, is still a highly relevant research direction according to leading scholars. (Conclusion 6)

5. CONCLUSIONS

We have drawn a number of conclusions from the literature on collective intelligence. The collective intelligence literature is a continuation of contributions in what has been called the "Information Age," a part of the "Digital Revolution." This is a development brought forward by natural and computer scientists, but where social scientists have a role to play, first by studying the applications and consequences that technologies have on people and societies. The body of literature published under the collective intelligence umbrella is truly interdisciplinary (C1). The association to the notion of wisdom of the crowds is problematic for several reasons. The journalist Surowiecki's idea is an erroneous interpretation of Galton's contribution about the regression towards the mean in statistics. Experience and empirical findings suggest instead that the wisdom of the knowledgeable is a more accurate term (C2). However, as our societies are becoming ever more short-sighted (financial markets, profits, product life cycles, etc.) there is an increasing gap between knowledge and wisdom in society. As a consequence, we argue that neither the crowd nor the knowledgeable are very wise (C3) and "wisdom of the wise" is a tautology and a meaningless expression. The content of the collective intelligence literature has been visited and revisited numerous times during the last half a century in the social sciences. As such, it can be seen as a part of a larger paradigm shift as noted in the first conclusion (C4). Just as with artificial intelligence, every revisit seems to bring something new and have great potential value. But, the collective intelligence literature strikes one not only by its lack of historical perspective, lack of good data in some of its leading publications, but by a general lack of critical sense as to the phenomenon studied (C5). Complex social systems seem still to be relevant for the study of intelligence related topics such as collective intelligence (C6). Stevan Dedijer made the same observations about the relation to social intelligence. The study of social systems based on evolutionary theory is a more fruitful scientific paradigm for the study of not only intelligence studies, but for the social sciences in general.

6. REFERENCES

Ackoff, R. L. (1971). Towards a system of systems concepts. *Management science*, *17*(11), 661-671.

Von Bertalanffy, L. (1968). General system theory. *New York*, *41973*, 40.

Boulding, K. E. (1956). General systems theory—the skeleton of science. *Management science*, *2*(3), 197-208.

Corvaja, A. S., Jeraj, B., & Borghoff, U. M. (2016). The Rise of Intelligence Studies: A Model for Germany? *Connections*, *15*(1), 79-106.

Dedijer, S. (2010). *Stevan Dedijer: My Life of Curiosity and Insights: a Chronicle of the 20th Century*. Nordic Academic Press.

Dedijer, S. (1999). Doing business in a changed world: The intelligence revolution and our planetary civilization. *Competitive Intelligence Review: Published in Cooperation with the Society of Competitive Intelligence Professionals*, *10*(3), 67-78.

Engel, D., Woolley, A. W., Jing, L. X., Chabris, C. F., & Malone, T. W. (2014). Reading the mind in the eyes or reading between the lines? Theory of mind predicts collective intelligence

equally well online and face-to-face. *PloS one*, *9*(12), e115212.

Franks, N. R., Pratt, S. C., Mallon, E. B., Britton, N. F., & Sumpter, D. J. (2002). Information flow, opinion polling and collective intelligence in house–hunting social insects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *357*(1427), 1567-1583.

Franks, N. R. (1989). Army ants: a collective intelligence. *Am. Sci. 77*, 138–145.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246-263.

Gregg, D. G. (2010). Designing for collective intelligence. *Communications of the ACM*, *53*(4), 134-138.

Hoogenboom, B. (2006). Grey intelligence. *Crime, Law and Social Change*, *45*(4-5), 373-381.

Hulnick, A. S. (2010). The dilemma of open sources intelligence: Is OSINT really intelligence?. In *The Oxford handbook of national security intelligence*.

Johnson, L. K. (Ed.). (2007). *Handbook of intelligence studies*. Routledge.

Keefe, P. R. (2010). Privatized spying: the emerging intelligence industry. In *The Oxford handbook of national security intelligence*.

Kuhn, T. S. (1962). The structure of scientific revolutions. *Chicago and London*. Lanier, J. (2006). Digital maoism. *The Edge. org*.

Lahneman, W. J. (2010). The need for a new intelligence paradigm. *International Journal of Intelligence and CounterIntelligence*, *23*(2), 201-225.

Leimeister, J. M. (2010). Collective intelligence. *Business & Information Systems Engineering*, *2*(4), 245-248.

Lykourentzou, I., Papadaki, K., Vergados, D. J., Polemi, D., & Loumos, V. (2010). CorpWiki: A self-regulating wiki to promote corporate collective intelligence through expert peer matching. *Information Sciences*, *180*(1), 18-38.

Lynn, R., & Hampson, S. (1986). The rise of national intelligence: Evidence from Britain, Japan and the USA. *Personality and individual differences*, *7*(1), 23-32.

Luhmann, N. (1968). *Zweckbegriff und Systemrationalität*. Tübingen: Mohr.

Luhmann, N., & Rechtswissenschaftler, S. (1984). *Soziale systeme: grundriss einer allgemeinen theorie* (Vol. 242). Frankfurt am Main: Suhrkamp.

Luttwak, E. N. (1990). From geopolitics to geo-economics: Logic of conflict, grammar of commerce. *The National Interest*, (20), 17-23.

MacKay, C. (1841). Extraordinary Popular Delusions and the Madness of Crowds. Amherst NY.

Magnusson, D. (Ed.) (1996). The lifespan development of individuals-behavioral, neurobiological and psychological perspectives: A synthesis.

Mainzer, K. (2007). *Thinking in complexity: The computational dynamics of matter, mind, and mankind*. Springer Science & Business Media.

Nguyen, N. T. (2008). Inconsistency of knowledge and collective intelligence. *Cybernetics and Systems: An International Journal*, *39*(6), 542-562.

Rheingold, H. (2002). Smart Mobs: the next social revolution Cambridge. *MA: Perseus*.

Rajasekhar, A., Lynn, N., Das, S., & Suganthan, P. N. (2017). Computing with the collective intelligence of honey bees–a survey. *Swarm and Evolutionary Computation*, *32*, 25-48.

Rathmell, A. (2002). Towards postmodern intelligence. *Intelligence and National Security*, *17*(3), 87-104.

Shum, S. B., Aberer, K., Schmidt, A., Bishop, S., Lukowicz, P., Anderson, S., ... & Edmonds, B. (2012). Towards a global participatory platform. *The European Physical Journal Special Topics*, *214*(1), 109-152.

Surowiecki J (2004) The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations. 1st Doubleday Books, New York

Sörensen, K. (2015). Metaheuristics—the metaphor exposed. *International Transactions in Operational Research*, *22*(1), 3-18.

Tammet, D. (2009). *Embracing the wide sky: A tour across the horizons of the mind*. Simon and Schuster.

Toffler, A., & Alvin, T. (1980). *The third wave* (Vol. 484). New York: Bantam books.

Wolf, M., Krause, J., Carney, P. A., Bogart, A., & Kurvers, R. H. (2015). Collective intelligence meets medical decision-making: the collective

outperforms the best radiologist. *PloS one*, *10*(8), e0134269.

Woolley, A. W., Aggarwal, I., & Malone, T. W. (2015). Collective intelligence and group performance. *Current Directions in Psychological Science*, *24*(6), 420-424.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *science*, *330*(6004), 686-688.

Young, M. (1958). *The rise of the meritocracy*. Routledge.

Zadeh, L. A. (2008). Toward human level machine intelligence-is it achievable? the need for a paradigm shift. *IEEE Computational Intelligence Magazine*, *3*(3), 11-22.

Journal of Intelligence Studies in Business
Vol. 9, No. 2 (2019) pp. 19-29
Open Access: Freely available at: https://ojs.hh.se/

Collective intelligence process to interpret weak signals and early warnings

Fernando C. de Almeida[a]* and Humbert Lesca[b]

[a]*University of São Paulo, Brazil*
[b]*Laboratoire CERAG UMR 5820 CNRS - Université de Grenoble, France*

Corresponding author (*): fcalmeida@usp.br

Received 10 October 2019 Accepted 15 October 2019

ABSTRACT The treatment of weak signals is identified as a method to identify strategic surprises in a firm's environment. Many researchers address the problem of anticipation of movements that have an impact on a firm's environment. Weak signals are considered in some approaches and presented in the literature, but also other methods are explored. This article tries to deepen the discussion of how to treat and interpret weak signals collected in a firm's environment. The concept of a weak signal is explained and the discussion about how to collect and interpret them is presented. Two important aspects are distinguished in the article: the usefulness of information technology in collection and treatment of weak signals and the concept of collective sensemaking in interpreting weak signals. Two cases of weak signal interpretation are presented as illustrations.

KEYWORDS Collective sensemaking, competitive intelligence, weak signals

1. INTRODUCTION

There is an ongoing lack of understanding of the notion of weak signals and few methods exist to explore them. Some researchers developed methodological procedures to explore them (Lesca and Lesca, 2014) and produced methods for collecting and interpreting weak signals in a competitive intelligence process. The traditional competitive intelligence process (Herring, 1988) considers key intelligence topics (KIT) to "provide the focus the prioritization needed to conduct effective intelligence operations and to produce the appropriate intelligence" (Herring, 1999, p.6). KITs are comprehended in the first step of planning and direction of competitive intelligence cycles. This step defines an organization's intelligence needs and orients the search of a firm in the competitive environment. Many organizations considerer the environment as analyzable and that it has the information needed to obtain correct answers to their questions. It is just a matter of searching for this information through "discovery", one of the four scanning methods that may be assumed by an organization in an environmental scanning process (Daft and Weick, 1984). There is no reflection and hypothesis of what may or may not exist in the environment. The information is there. The intelligence needs and KITs identified in the "planning and direction" step, conducted in the competitive intelligence search process.

Some organizations may consider the environment unanalyzable and adopt "enacting" as a strategy to approach the interpretation of the environment. "The organization in some extent may create the external environment. The key is to construct, coerce, or enact a reasonable interpretation that makes previous actions sensible and suggests some steps. The interpretation may

shape the environment more than the environment shapes the interpretation" (Daft and Weick, 1984, p.287).

Weak signals suggest interpretation and sensemaking (Shoemaker and Day, 2009) as the environment is considered unanalyzable. Weak signals through an inductive process stimulate the hypothesis generation and sensemaking of the competitive environment. It is not a matter of finding the right answer, as in a discovery scanning processes, but creating perspectives and possibilities that outline how the future environment and competitive may move (Gilad, 2011). Whatever future is considered, the future does not exist yet and the perspective of today may not happen in the future. For example, possible competitive moves identified today may not happen if the environmental scenario changes as a consequence of economic change, competitive moves or any other unexpected environmental change.

In this article Ansoff 's concept of weak signal and an operational process of treating this weak signal is discussed. It allows us to create hypotheses and perspectives about future moves in a competitive environment that may impact an organization. Weak signals are considered here to be an inductor of collective sensemaking about what may or may not come in a future environment.

2. THEORETICAL BACKGROUND

2.1 Weak signal

Information of an anticipatory nature is a weak signal. The notion of weak signals, a type of metaphor (Ansoff, 1975), has proven interesting on account of its orientation toward attention given to surprises and ruptures that may occur in the business environment. "For the first time, the idea of a need to be 'early', or rather as early as possible in anticipating change, was expressed and translated into a complete methodological proposal" (Rossel, 2012, p.230).

However, the weak signal definition lacks precision and does not constitute actionable knowledge (Argyris 1996), despite the fact that Ansoff (1975) clearly attributes an anticipatory character to weak signals. According to the author, these fragments of information have a propensity to trigger, in the entrepreneurs that observe them (provided they pay attention), a sensation that something important may happen in the general environment. "For Ansoff, any change taking place is preceded by some form of 'warning', which the analyst has the role of capturing and making good use of. This is what he called a signal, based on the Information Theory work of Shannon and Weaver in the 1940s" (Rossel, 2012, p.230).

This sensation approaches that of intuition, triggered by data that is perceived and then examined attentively. Such information plays a triggering role, inducing the stimulus of an interrogation followed by an interpretation. Next, an inquiring entrepreneur will wish to know more about the question and obtain further information to refine this sensation. Before the interpretation through a weak signal, the decision maker had probably not asked for anything concerning the subject as his/her attention was not activated.

This notion of a weak signal does not have an operational definition. In practice it can be seen that expressed weak signals are misinterpreted in companies and generate contradictions (Lesca, 2011).

2.1.1 Meaning of the word weak: contradiction and propositions

Our experience through numerous company-based action research projects leads us to verify that the expression of weak signals is misinterpreted by most entrepreneurs due to the adjective *weak*. We often hear: "We don't want to capture weak signals, but strong ones!" Evidently, the word "weak" leads entrepreneurs in the wrong direction. Indeed, a signal can be weak in its appearance and thus discrete in terms of meaning but potentially very rich in meaning; in this sense it can "announce" something very important for the individual that is able to capture and interpret it.

In our view, Ansoff (1975) meant that a signal can be classified as "weak" if it bears the following characteristics:

a) Fragmented: for example, there is only a fragment of information from which it can be attempted making inferences in a holistic procedure. It is expect that the number of weak signals is very small. It is not a context where the amount of information is high and it is not a matter of treating a huge amount of data.

b) Submerged amidst myriad bolder data: it is weak because it is submerged, mixed with a myriad of useless information that creates noise. It appears thus, with weak visibility,

most people pass over these signals, barely noticing them.

c) Meaning not evident: it is weak because of an apparent weak meaning and ambiguousness. Information such as a weak signal does not bring a visible interest. On the contrary it is equivocal or ambivalent. This information is of little significance by itself, and does not have an evident connection with other information.

d) Unexpected, not familiar, non-repetitive, and it risks not to be noticed: The concept of non-familiarity of this kind of information makes it difficult to distinguish. Cognitive biases may also distort its identification or interpretation and analysis in competitive intelligence processes (Memheld, 2014).

e) The operational utility of a weak signal is not immediately evident and it seems not to be very useful. The very same information may be of importance for one person and of no interest for another. It is not evidently interesting, the consequences of the event identified are not evident.

f) The detection of a weak signal is difficult. Because of this the opportunity to use information technology or big data techniques to search for weak signals on the web, or on a newspaper's site is high (Lesca, Buitrago and Casagrande, 2016, Buitrago, Casagrande and Lesca, 2015). The technology can select news with potential weak signals to be evaluated.

Nevertheless, weak signals are at the core of anticipatory, strategic intelligence because they are of potential use to managers, if the managers are able to perceive and interpret them. This type of information can range from indicators of disruptions (Ansoff, 1975) to larger events, and they clarify the intentions of external actors (competitors, clients, suppliers, and various signs of changes in general).

Individual differences may also influence the interpretation and importance perceived of information (Stanovich and West, 2012).

2.1.2 Definition of a weak signal

As posited by Ansoff (1975), a weak signal is a "datum," often with an insignificant appearance and submerged in myriad other data, the interpretation of which can warn that an event (perhaps not yet initiated) is about to occur and is likely to have significant consequences in terms of risks or opportunities. It has an anticipatory feature (Lesca 2003). Weak signals have the following characteristics presented by Lesca (2001):

- Fragmented: To which information can it be related?
- Isolated
- Uncertain reliability: Is it possible to relate it to something else?
- Imprecise
- Unpredictable: Where to look, when to pay attention to the information?
- Ambiguous
- Apparently little or no utility: How to avoid ignoring it?
- Anticipatory
- No standardized key words: How to access it?
- Unusual, singular, unfamiliar: When to pay attention to it?
- Possibly intentional on the part of the signaler
- Submerged amidst a large quantity of data: How to notice it?
- Subjective
- Often qualitative

2.1.3 Characteristics of a weak signal (adapted from Lesca 2001)

Weak signals originate from two types of sources. Contacts with the field: personal relationships, visual observations, etc. These are the richest sources of anticipatory information, but significant human aptitudes are needed to exploit them. Databases, the internet, websites, etc. These sources have been causes of data overload (Edmunds and Morris 2000; Lesca et. al. 2009, Sherkock, 2011). Lately, efforts at using new technologies are helping to deal with large data sources on the web to identify weak signals (Buitrago-Uitrago, 2014, Casagrande, 2012), and to limit the information overload (Lau et al., 2012).

One should not use anticipation and prediction interchangeably. Prediction is mainly the calculation of the trends in the quantitative database collected over a period. The calculation does not include singletons or outliers, and computers are of great use. It is more related to Daft and Weick's (1984) discovery processes. Prediction may be

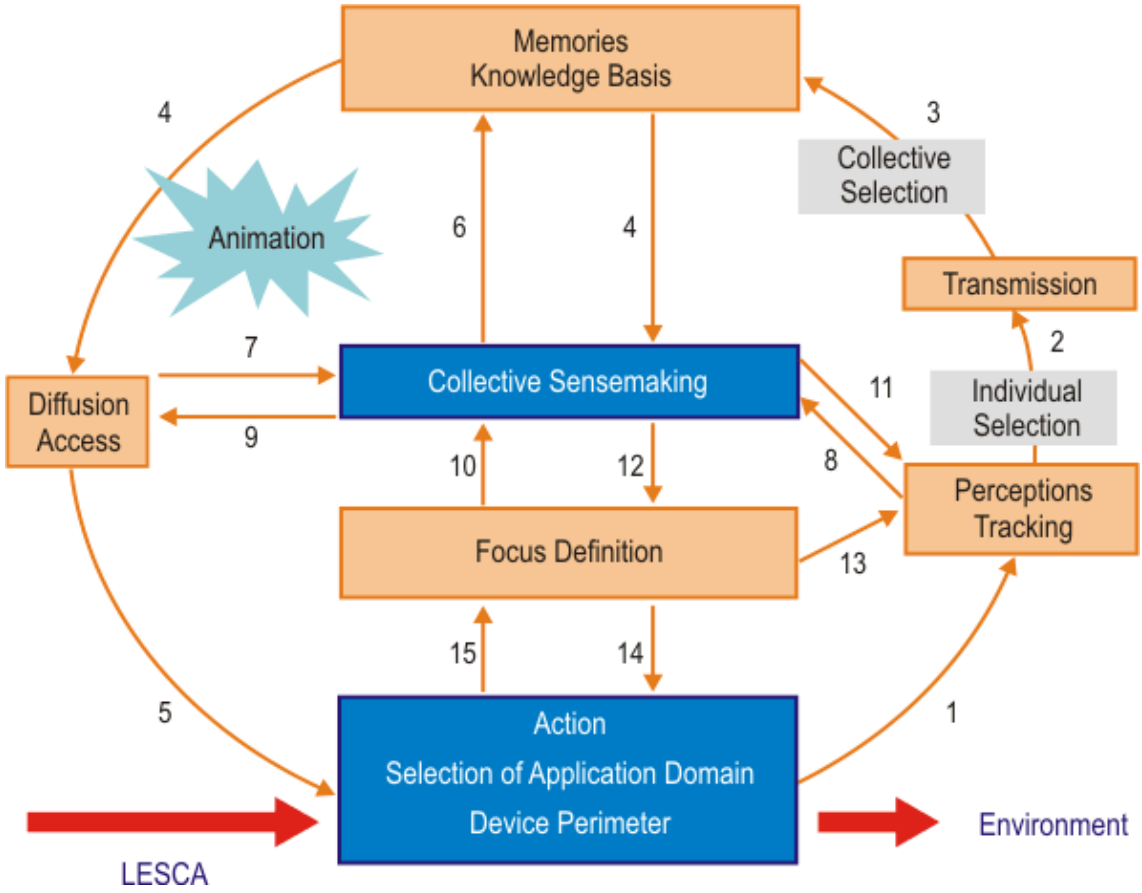


Figure 1 SCIS Conceptual Model.

expressed by a curve integrating a significant part of the data (for example, 80% of the observations), extrapolating to the future what was learned from the past. The 20% of the observations not integrated in the predicting curve are considered to be less important or outliers.

Anticipation concentrates singular information or outliers left aside by prediction-makers. It is interested in the 20% of the data left aside by the predictions. Though considered outliers by the statistics, this is possibly where weak signals can be detected early, as well as possible surprises, discontinuities or disruptions. These weak signals should be stimuli for strategic management (Reinhardt,1984; Starbuck and Milliken, 1988; Gilad, 2004; Marrs 2005).

Consequence 1: The first question to be presented is: "What is one's objective: to predict or to anticipate?" If it is to detect surprises, ruptures, or breakthroughs, then weak signal treatment is a appropriate method.

Consequence 2: Information like weak signals are the one considered by a process that Daft and Weik (1984) called enacting, where a process of sensemaking and interpretation is induced by the weak signal.

The treatment of weak signals stems from interpretation through collective sensemaking, and not an algorithm with information technology (Daft and Weick 1984).

2.2 Detection/ acquisition of weak signals

Strategic Scanning Information Systems (SCIS) is the way by which a firm seeks to detect signals as early as possible, before the occurrence of changes in the environment, so as to secure sustained competitiveness. It is a collective, transversal, proactive, and continual process through which a group of individuals collaborate to pursue, capture, and use information of an anticipatory nature concerning the external environment and changes that can be produced there (strategic surprise), including disruptions (Lesca, 2003, p10). A conceptual SCIS model is shown in Figure 1.

Over successive experiments in different organizations, it was possible to distinguish two types of strategic monitoring processes. There are those including a phase of collective sensemaking that is particularly important in recognizing and exploring weak signals. It can be referred to as an anticipatory collective intelligence processes in the sense of Daft and Weick's enacting process. The second is those that do not include the collective sensemaking phase. This type of process is currently the most-used by companies. Daft and Weick called it a discovery process.

According to Daft and Weick (1984, p291), there are four different ways to interpret the environment, leading to four different modes of organizing processes for scanning the environment (four quadrants). Daft and Weick (1984) suggested a model to categorize organizations according to the way top management interprets the collected information to make a decision and to define actions. They suggested the existence of a relation between strategic orientation and the way firms monitor the environment, based on Aguilar's (1967) and Miles and Snow's (1978) models respectively. Daft and Weick (1984) used two dimensions to explain how organizations approach environmental knowledge. The first one is how much top management considers the environment stable and the second one is how actively the organization searches information allocating resources. From these dimensions, four ways of interpreting the environment are derived: undirected viewing for reactive organizations that obtain information randomly; conditioned viewing for defensive organizations that frequently use information that once in the past was helpful; discovery for analytical organizations that intend to formally search and structure environmental knowledge; enacting for prospective organizations that intend to transform the environment through innovation and are characterized by informal searches of information.

Daft and Weick (1984) suggest two different dimensions concerning the scanning strategy and firms environment perception. The first is intrusiveness. The firm exhibits proactive behavior, searching for business opportunities, and strives to prevent all types of threats. To this end various sources of information are accessed (formal and informal, documented and field-based). It seeks several types of information (field-based, formal, and digital information). People in charge of collecting information belong to different parts of the organization. Exploiting the information, mainly weak signals, is done through interpretative processes, considering the characteristics of weak signals presented above. The results of the interpretations aim to assist in strategic decision-making.

The second is the unanalyzable dimension. The enterprise is in an unanalyzable environment. The sources of information are diverse, but the richest are also the least formal: human contact is essential. Information collection is not done by a bureaucratic "cell," but is entrusted to collaborators with main activities other than scanning. Perception processes are essential. Exploring information is not automated: it is mainly based on human and heuristic cognitive processes. People interpret information individually and then collectively. Collective learning is important. Understanding weak signals advances by trial and error, or "learning by doing".

2.3 A collective sensemaking of weak signals

Weak signals are of little interest per se. They start to become useful if it is known how to exploit them to create a useful meaning for strategic management (Haeckel 2004). The treatment of weak signals lies in the resulting interpretation (Daft and Weick 1984). Information technology is becoming more and more effective in detecting weak signals automatically (Lesca, Buitrago and Casagrande, 2016, Buitrago, Casagrande and Lesca, 2015). However, interpretation can only be made by individuals, alone or in groups (Almeida, 2009), as interpretation is also a matter of a decision maker's perspective (Gilad, 2011).

It was shown that the characteristics of weak signals create a number of difficulties when considering its features. Lesca (1995) suggest that the exploitation of weak signals could be accomplished with heuristics. The conceptual model for the application of the heuristics was illustrated in Figure 1 and agrees with the works of Daft and Weick (1984) and Nonaka (1991, 1994). Lesca and Lesca (2014) suggest that heuristics must be used within a collective working group of people chosen according to their involvement in the subject and their knowledge. The work of collective interpretation is called "collective intelligence" (Lesca and Caron 1996; Blanco

and Lesca 1998; Blanco et al. 2003; Lesca, 2003).

Collective sensemaking is the operation of collective interpretation thanks to which meaning and knowledge are created from weak signals (input) that have the role of inducing stimuli, and through interactions among participants (Mamykin, Nakikj and Elhadad, 2015, Lesca, 1995). The result of collective sensemaking (output) is the formulation of plausible future views capable of orienting entrepreneurs (Lesca and Caron 1996). The collective sensemaking accomplished according to Lesca's (1995) heuristics is in line with Schoemaker and Day (2009). Collective sensemaking cannot be understood as "organizational sensemaking," because experience shows us that it is not possible to mobilize all people within a firm to interpret information.

The process of collective intelligence arises from a group of individuals when the signals coming from the competitive environment are collected, selected, interpreted, and compared through collective work so as to make sense. It is a process in which group members interact in different ways, subject to behavior rules of collective work (Lesca, 2003). A weak signal must be examined from different points of view, by different people holding different positions within a firm (Starbuck and Milliken 1988).

The discussion of collective sensemaking appears in the academic literature in different domains like teaching (Coburn, 2001), on-line services (Mamykin, Nakikj and Elhadad, 2015) and competitive intelligence (Soilen, 2017, Lesca, 1995).

The discussion of weak signals in a collective way is in line with the idea that in a competitive intelligence process, it is not effective to deliver reports and answers to managers as they tend to ignore them or to consider them threatening to their position (Soilen, 2017). In a collective process around a group of individuals, they debate and discuss the environment. The role of the competitive intelligence staff is to conduct and help the discussion process. Decision makers then may have insights about the market, have their own perspectives about what is going on and take decisions based on their own perspectives (Gilad, 2011, Rohrbeck and Bade, 2012).

3. TWO CASES OF WEAK SIGNAL INTERPRETATION AND SENSEMAKING

Two examples are explored here to access the concept of weak signals as follows. These two examples were treated by the team involved in this research in order to analyze weak signals for to companies.

The first is the ABB Case (Lesca, H., Buitrago-Uitrago A. F. ,Casagrande, A., 2015).

Let's consider a company with a strategic intelligence process that is interested in ABB as a target of the process. The information to be treated in the following paragraphs was presented as follows:

> "ABB wins the Energy Prize at the Arabian United Emirates."

Why can this data be considered a weak signal? It is fragmented (less than a line). It was taken from a newspaper that contains over thirty pages per day. It is submerged in a huge volume of data.

How can this be seen a warning sign in this weak signal?

- Pertinence. Considering ABB as an example of a target, this is a fragmented piece of information.
- Surprise. This data was not expected, caught someone's attention, and triggered a process of collective reflection. As of that moment, this data gained the status of information for us.
- Importance. Considering ABB as a target and the motivations justifying a process of strategic intelligence, it can be raised the hypothesis that ABB relations could suggest business opportunities for the company interested in it. A manager considering this information observed: "The information thus began to be potentially useful to us. We could enter the Arabian market through ABB."
- Anticipation. Is this information anticipative? It is clear that ABB prize is already a past event. On the other hand, it could be estimated that there still may be initiatives not known of ABB in Saudi Arabia showing opportunities.

The set of collective reflections by a group in the company dealing with the information, led it to see in the weak signal as a warning sign. Thus, it was possible to exploit a weak signal and trigger the concrete action of contacting ABB. This procedure gives rise to a positive

output beyond the company's initial expectations. This example shows that, in certain cases, detecting weak signals and transforming them into early warning signals fully exploited by the firm's leadership generates benefits that may be far superior to costs.

The second example is the AZULY case. The information was presented as follows:

"P. AZULY goes to the X group".

Why consider this data to be a weak signal? It is very fragmented, qualitative data. At first it was captured through oral communication, talking with a work associate. Later it was found printed in a recent issue of a professional magazine. The information occupied only two lines—a piece of news submerged in a 150-page magazine, bound to go unnoticed. The utility of this information leading to action was not evident. Furthermore, this data is ambiguous and open to multiple interpretations. It was a surprise. It caught the researcher's attention almost by chance. The piece of information started to have a meaning for the team.

The information is probably anticipative: the strategic operation of the X group is only in the initial stage of its preparation. A field expert that was contacted informs us that this sort of operation and a communication campaign related to the strategic topic possibly identified requires around 12 months of preparation.

In conclusion, in this example one moves from a weak signal to an early warning signal (Gilad, 2003). Clearly, the latter is based on hypotheses (Lesca, 2014) that are formulated and are able to be verified. Such interpretation of the weak signal is not the only one possible. It allows the decision maker to be placed in an "alert mode". From then on, it is up to him/her to accomplish what is necessary to further explore the situation and reduce the uncertainty if it is judged useful.

But what type of usefulness does this weak signal represent to the X group? The strategic operation was revealed to be of great importance, both for the X and Y groups. Group Y had available to itself of a sufficiently long term of anticipation to create plans to consider an offensive vis-à-vis X.

4. A STRATEGIC INTELLIGENCE METHOD

In order to organize the detection, capture, and exploration of weak signals, Lesca (2003) suggests the LEScanning (Learning Environmental Scanning) method. Figure 1 indicates the different blocks that make up the entire process of anticipatory strategic intelligence.

4.1 Domain delimitation

Approaching the SCIS (Strategic Scanning Information System) device: a company can have various SCIS devices. In a large company, for instance, there are devices at the company level, together with the CEO, or at the group level when the company comprises a number of autonomous units or "business units."

Perimeter delimitation of the SCIS device: perimeter refers to the list of people included in the device, each of whom will have to contribute and will experience some benefit.

4.2 SCIS target

Targeting is the operation of delimiting the portion of the environment-of-interest to the members of the perimeter of the future SCIS device. Focusing means expressing in an explicit and formal manner who/what can serve as a common interest for the different participants of the SCIS process.

4.3 Collecting/surrounding the information by designated people

This phase requires human and formative qualities. It is an elementary form of the perception filter (Starbuck and Milliken 1988).

4.4 Information selection

This consists of retaining, from the collected information, only that which is of interest to potential users within the SCIS perimeter. This is a crucial operation: lack of selection leads to data overload and suffocates the SCIS process, whereas too restrictive a selection impoverishes and empties the SCIS process.

Selection (or filtering) is the separation of raw data from potentially weak signals. It is conducted by taking the target into account.

4.5 Collective sensemaking

This is the process of exploring weak signals to create sense. The interpretation of weak signals cannot be valid if conducted by just a single person. It requires plurality and competing viewpoints from people with different knowledge, experience and points of

view. But it requires a certain familiarity with

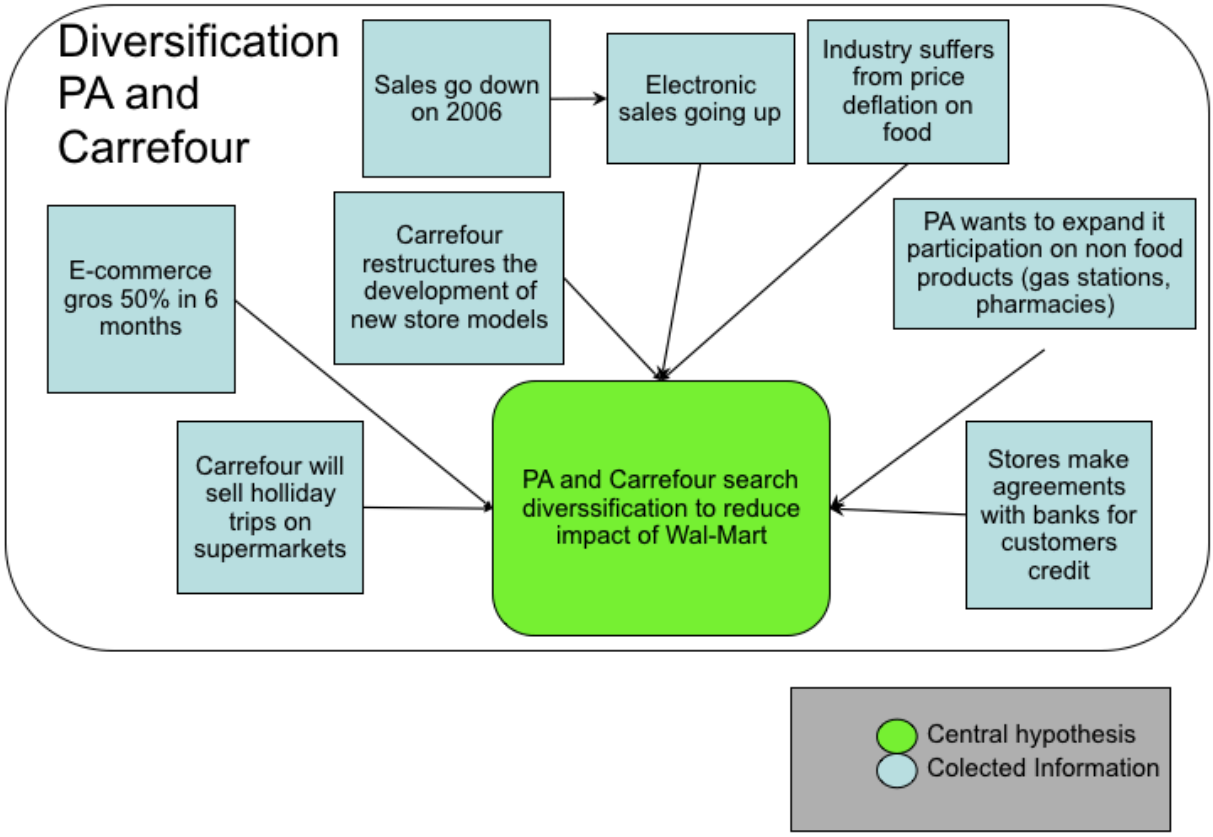


Figure 2 Example of puzzle: the Carrefour case.

the subject. Interactions among people are very important.

It can be suggested that heuristics creating links between the pieces of information (weak signals) used during the collective work session can map fragments of isolated information into a more significant and reasoned visual (or other) representation. Figure 2 shows an example of a puzzle, referring to the Carrefour example in Brazil.

The collective interpretation of weak signals may imply resorting to a single or several external specialists. Lesca and Kriaa (2007) conceived and tested a method of remote monitoring to help the leader of the collective sensemaking sessions using the Puzzle method.

5. CONCLUSION

Ansoff (1975) distinguished the importance of treating weak signals to identify strategic surprises. The point was to identify disruptions and strategic surprises, not tendencies projected from past data. His article comes after some decades of a stable environment and continuous growth where long-range planning was still possible. However, the stable environment from the 1950's and 1960's changed and the environment became turbulent and the experience and projections from the past were not enough to anticipate the future. Formal search is questionable in its ability to predict the future, since it is strongly associated with analyses and statistical predictions that may divert the attention from strategic surprises or disruptions (Ansoff, 1975). Data from the past may be interesting to identify future outcomes only in a stable environment. In this case quantitative data analysis may be of use. Ansoff suggested the importance of paying attention to weak signals that might preannounce changes in the future environment. Kahaner (1997), sharing the same reasoning, comments that one of the most difficult tasks of monitoring the competitive environment is to predict what will happen in the future and that quantitative information, in general, describes the past and therefore suggests that even unstructured information such as rumors and comments should also be part of the scope of monitoring. Rumors may be weak signals of future events.

Decades after Ansoff's proposition, the discussion about weak signals and early warning was extended. Different authors

reinforced Ansoff's preoccupation with this kind of information. They proposed useful approaches to increase firms' attention to not so clear events that might suggest important moves in the environment. Rossel (2012) identified different "neo-ansoffian contributions" (p.232), considering them diverse and rich. The author considered classificatory maps the richest one, for example, where Morrison and Wilson (1996) made cross references to probabilities of occurrence with impact concerning weak signals. This kind of approach is particularly interesting as it suggests ways of interpreting weak signals. Day and Schoemaker (2005) proposed to scan the periphery in order to identify events not in the main stream of the decision maker's attention.

Treating weak signals requires methods that enable one to identify and interpret them. Because the characteristics of weak signals make them difficult to be identified and interpreted, there is still a considerable opportunity concerning new ways of working on them.

The present study intends to bring some methodological propositions and suggestions. One important aspect of treating weak signals to be further explored is the use of information technology. It may help in identifying and treating weak signal interpretation. It also requires intuition, imagination, and sensitivity in their interpretation, a task that cannot be fully accomplished by information technology, though it is increasingly helpful in the first steps of collection and interpretation of weak signals. It is also importance to distinguish the collective reflection on the eventual meaning of the weak signals, as different persons bring different knowledge and perspective to a discussion.

As suggested in the present paper, the most important support that can be brought by the strategic intelligence processes lies in anticipating surprises and ruptures.

Our experience shows us that weak signal treatment enables long term visibility and enhanced anticipation of threats and strategic opportunities in the environment. The treatment of weak signals requires us to consider their characteristics.

6. REFERENCES

Almeida, F.C. de (2009). People – A Critical Success Factor in the Brazilian CI Process. Competitive Intelligence. Vol. 12, n.3, May-June, p. 13-19. Accessible sur http://www.advsbrasil.com.br

Ansoff, H.I. (1975) – Managing strategic surprise by response to weak signals. *California Management Review*, vol. XVIII, n°2, p.21-33

Akhavan, P., and Pezeshkan, A. (2014). Knowledge management critical failure factors: a multi-case study. VINE,Journal of Information and Knowledge Management Systems. 44(1), 22–41. doi:10.1108/vine-08-2012-0034

Argyris, C. (1996,) "Actionable knowledge: Design causality in the service of consequential theory", *The Journal of applied behavioral science*, vol. 32, nº 4, pp. 390-406.

Blanco, S. Lesca, H. (1998) – Business Intelligence : a collective learning proces\for the selection of early warning signals. ECIS' Workshop on Knowledge Management, Aix-en- Provence, juin, 12 p. Edité par CEA/DIST, Saclay, France, 15p

Blanco, S. Lesca, N. (2003) - From weak signals to anticipative information: learning from the implementation of an information selection method. In Search of Time: proceedings of the international conference: Isida - Palermo, Italy, May 8-10, 2003. – Palermo : F.Orlando, 2005. pp.197-209.

Buitrago-Uitrago A. F. (2014). Aide à la prise de décision stratégique. Détection d'Informations pertinentes de sources numériques d'Internet. Spécialité : Sciences de Gestion. [s.l.] : Université de Grenoble Alpes.

Casagrande, A. Proposition d'une mesure de voisinage entre textes : Application à la veille stratégique. Spécialité : Mathématiques-Informatique. [s.l.] : Université de Grenoble, 2012.

Coburn, C.E. (2001). Collective Sensemaking about Reading: How Teachers Mediate Reading Policy in Their Professional Communities. *Educational Evaluation and Policy Analysis 23*, 2, 145–170.

Lesca, H., Buitrago-Uitrago A. F., Casagrande, A. (2016). Signaux 'faibles' anticipatifs (SFA) pour la prise de décisions stratégiques : deux outils pour les détecter et les utiliser. 13e Forum Européen, IES, Rouen.

Daft, R. L., Weick, K.E., (1984) – Toward a model of organizations as interpretation systems. *Academy of Management Review*, vol.9, n°2, pp.284-295

Edmunds, A., Morris, A. (2000) – The problem of information overload in business organisations: a review of the literature. International Journal of Information Management. v. 20, p.17-28.

Day, G. S., and Schoemaker, P. J. H. (2005). Scanning the periphery. Harvard Business Review, 83, 144–148.

Gilad, B. (2004) - *Early Warning: Using Competitive Intelligence to Anticipate Market Shifts, Control Risk and Create Powerful Strategies*. Ed. Amacom.

Gilad, B. (2011). Strategy without intelligence, intelligence without strategy. Business Strategy Series, 12(1), 4-11.

Herring, J. P. (1988). *Building a Business Intelligence System. Journal of Business Strategy, 9(3), 4–9.* doi:10.1108/eb039219

Herring, J. P. (1999). *Key intelligence topics: A process to identify and define intelligence needs. Competitive Intelligence Review, 10(2), 4–14.*

Lau, R. Y. K., Liao, S. S. Y., Wong, K. F., Chiud, D. K. W. (2012). Web 2.0 environmental scanning and adaptive decision support for business mergers and acquisitions. *MIS Quarterly.* december, 36,(4)1239-1268.

Lesca, H. (1995) - *The crucial problem of the strategic probe: the construction of the 'Puzzle'.* CERAG – CNRS-5820, série Recherche 95-02, avril, 24 p. accessible www.veille-strategique.org.

Lesca, H. (2001) – *Veille stratégique : passage de la notion de signal faible à la notion de signe d'alerte précoce*. Colloque VSST 2001, Barcelone oct., Actes du colloque, tome 1, pp. 98-105. Accessible www.veille-strategique.org

Lesca. H (2003) veille stratégique : la méthode L.E.SCAnning ®, Editions EMS. 180 p, accessible www.veille-strategique.org

Lesca, H., Kriaa, S., Casagrande, A. (2009) - *Veille stratégique : Un Facteur d'échec paradoxal largement avéré : la surinformation causée par l'Internet. Cas concrets, retours d'expérience et piste de solutions*. VSST 2009, Nancy accessible www.veille-strategique.org

Lesca, N., Caron-Fasan, M-L (2008) - Strategic scanning project failure and abandonment factors: lessons learned. *European Journal of Information Systems,*17, 371-386.

Lesca, H., Lesca, N. (2014). Weak Signals for Stratetegic Intelligence - Anticipation Tool for Managers. ISTE Wiley.

Lesca, N., Caron-Fasan, M.-L., Aguirre, E.L. and Chalus-Sauvannet, M.C. (2016). Drivers and barriers to pre-adoption of strategic scanning information systems in the context of sustainable supply chain. Systèmes d'information and management 2015/3. 20, 9-46. DOI 10.3917/sim.153.0009

Marrs, R. (2005) – Early warning signals, A Convertion for Exploration – Part 1. Coemergence. Accessible on Internet.

Memheld, P. (2014). Intelligence analysis and cognitive biases: an illustrative case study, *Journal of Intelligence Studies in Business,* 4(2), 41-50.

Mamykina, L., Nakikj, D., and Elhadad, N. (2015). *Collective Sensemaking in Online Health Forums. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15.*

Morrison, I. Wilson (1996) The strategic management response to the challenge of global change, in: H. Didsbury (Ed.), Future Vision, Ideas, Insights, and Strategies, The World Future Society, Bethesda, MD.

Nonaka, I. (1991) - "The Knowledge-Creating Company..." *Harvard Business Review* 69(6), 96 - 104.

Nonaka, I. (1994) - A dynamic theory of organizational knowledge creation. *Organization Science*, 5(1) p.14-37.

Reinhardt, W.A (1984). An early warning system for strategic planning. Long Range Planning ,vol.17, n°5, p.25-34.

Rohrbeck, R. and M. Bade (2012). Environmental scanning, futures research, strategic foresight and organizational future orientation: a review, integration, and future research directions, ISPIM Annual Conference, Barcelona, Spain.

Rossel, P. (2012). Early detection, warnings, weak signals and seeds of change: A turbulent domain of futures studies, Futures, 44(3), 229–239.

Søilen, K.S. (2017) Why care about competitive intelligence and market intelligence? The case of Ericsson and the Swedish Cellulose Company. *Journal of Intelligence Studies in Business*, 7(2), 27-39.

Sherlock, A. (2011). Managing information overload. *Pharmaceutical Technology Europe, 23(8),* 12-13.

Shoemaker, P. J. H. And Day, G. S. 2009. How to make sense of weak signals. MIT Sloan Management Review, 50(3): 81– 89.

Stanovich, K and West, R. (1998). Individual differences in rational thoughts. *Journal of Experimental Psichology, 127(2)*, 161-188.

Starbuck, W.H. Milliken, F.J. (1988) - Executives' perceptual filters: what they notice and how they make sense. In D. Hambrick (Ed.), *The executive effect: concepts and methods for studying top managers*, Greenwich, CT: JAI Press, pp.35-65.

Wenger, E. *Communities of Practice: Learning, Meaning and Identity*. Cambridge University Press, Cambridge, 1998.

Weick, K.E., Sutcliffe, K.M., and Obstfeld, D. Organizing and the Process of Sensemaking. *Organization Science 16*, 4 (2005), 409–421.

Journal of Intelligence Studies in Business
Vol. 9, No. 2 (2019) pp. 30-42
Open Access: Freely available at: https://ojs.hh.se/

Study on the various intellectual property management strategies used and implemented by ICT firms for business intelligence

Shabib-Ahmed Shaikh[a]* and Tarun Kumar Singhal[b]

[a]*Symbiosis International (Deemed University) (SIU), Pune, Maharashtra, India*
[b]*Symbiosis Centre for Management Studies (SCMS), Symbiosis International (Deemed University) (SIU), Noida, Uttar Pradesh, India*

Corresponding author (*): shabib.ahmed@gmail.com

Received 30 September 2019 Accepted 25 October 2019

ABSTRACT Software technology is seeing enormous growth as it is used in all fields of technology. It is continuously evolving at a rapid pace and has a short span of the technological life cycle. The use of the software is not restricted only to information and communication technology but is used in all fields of technology. In many cases, the inventive step of a product or service lies solely in the software. Hence, the software plays a crucial role in all fields of technology. However, ease of copying poses a financial risk for the software industry, thereby creating major disincentives to the development of innovation. Still, the technology is changing very fast and firms investing in this technology expect quick returns on their innovation investments. Strategies for generating and managing intellectual property have subsequently taken center stage for information and communication technology companies, and patents have become an important feature providing maximum protection for any technology. Hence, intellectual property rights strategies in general and patenting strategies especially play a crucial role in the information and communication technology industry to be globally competitive. Firms never publish or disclose their intellectual property strategies; hence, this study makes use of the literature review to highlight various intellectual property management strategies used by information and communication technology firms for managing their intellectual property. These strategies can be offensive or defensive and may be used as proactive or reactive depending on various aspects such as market, territory, technology, or time. The insights provided in this work may help the research community from the IT domain in industry and academia to learn and modify their strategies for patent acquisition.

KEYWORDS Business intelligence, competitive intelligence, IP strategies, organizational performance, patents

1. INTRODUCTION

1.1 Information Technology

Information and communications technology (ICT) is often used as an extended synonym for information technology (IT). IT is the application of computers and telecommunications equipment to store, retrieve, transmit, and manipulate data, often in the context of a business or other enterprise. IT encompasses the inputting, storing, retrieving, transmitting, and managing data through the use of computers and various other networks, hardware, software, electronics, and telecommunication equipment (IPO, 2013). The core elements in the application of IT are

computers and their peripherals consisting of hardware and software.

1.2 Intellectual Property

Intellectual property (IP) is an intangible asset created from a human mind and having some value (Kavida & Sivakoumar, 2008; Isa et al., 2009). Intellectual property rights are the rights conferred on the persons for exploiting their intellectual property within a specified territory for a specific period. The intellectual property rights framework provides various alternatives for protecting the intellectual property generated from a business or required for a business to be globally competitive (WIPO-b). The exploitation and management of this intellectual property is often linked with business sales, export quality and marketing needs, along with research direction strategies to ensure that a firm remains competitive in a business (Zhang & Yang, 2016; Mahajan et al., 2015; Debackere & Veugelers, 2005; Zahra & Nielsen, 2002; Torvinen, & Väätänen, 2014). The full value of IP can be perceived as an information source derived from its technical details available in patent data, its uniqueness, and its volume as over 100 million patent documents that are freely available online for use as early as 18 months after the filing of a technology (Khode & Jambholkar, 2017). Parr and Smith (2016) point out that the commercialization of IP involves annual revenues of at least 5 trillion USD. Managing IP in general and patents in particular, has thus become crucial for the IT industry to survive. It is continuously evolving, has a short technological lifecycle, and is hit by many legal challenges towards its protection, litigations, and trolls (Shaikh & Londhe, 2016).

1.3 Strategies

Strategies are futuristic plans conceived before execution, depending on a set of predefined rules or previous experiences. Krig and Sandra (2017) define strategy as "the determination of the basic long term goals and objectives of an enterprise, and the adoption of courses of action and the allocation of resources necessary for carrying out these goals." The main aim of strategies is to sustain long term competitive advantage in business via means of building defenses against competitive forces (Porter 1993). Strategies can be proactively planned or reactive, based on situations and market places.

1.4 The need for IP strategies in IT

The IT industry has rapidly globalized (Cameron et al., 2006). As the software market started from the US, the US acts as a trendsetter for the protection of software via patenting. Other countries follow the US in protecting software via patents (Cameron et al., 2006) as this protection promotes a nation's technological innovation (Wang et al., 2012). A fundamental problem for the software industry is the ease of copying, which often poses a financial risk (Rao, 2001). This even creates significant disincentives to the development of new and innovative software programs, hindering software development (McGowan et al., 2007). Robust R&D operations are undertaken if protection is provided, which leads to the start of profitable businesses. Failure to protect software firms' developed products might affect a company's ability to operate freely at the primary level in the global market (Clarkson & Dekorte, 2006), which in turn would threaten a firm's own existence (Dedrick & Kraemer, 1993; Jyoti et al., 2010). Software innovations are usually incremental, fast-changing, and have a short lifecycle. Software is becoming more complex and sophisticated daily, with value-added features. Firms investing in this continually evolving and changing technology expect concrete protection for their IP and quick returns on their investments (Shaikh & Londhe, 2016).

In the field of information technology, trade secrets, copyrights, and patents are mainly considered for protection. While each of these has its advantages and disadvantages, patents are considered to provide the highest protection in the ICT sector, specifically for software (Shaikh & Londhe, 2016). Patents qualify the protection of the functional aspect of a product, process, or service, along with its underlying idea. The idea behind this is that software can easily be copied and independently developed when it comes into the market, and hence trade secrets, as well as copyrights, prove to be weak in protection. Additionally, copyrights are meant to protect the nonfunctional aspects and expression of ideas and not the functional aspects and ideas. Hence patent protection in the field of IT and mainly for software is gaining importance. At the same time, protecting software under patents also ensures that no one company can claim a monopoly under a particular innovation, thereby increasing competition (OECD, 2008; the United States. Federal Trade Commission, 2003). Many important

innovations have reached the marketplace with the help of the patent system (EPO, 2013). Different patent filing strategies are used by firms to gain a competitive advantage and survive and thrive in the market place (Shaikh & Singhal, 2018). This study focuses on patenting strategies of IT firms and uses it interchangeably with the term IP strategy.

2. IP STRATEGIES FOR BUSINESS INTELLIGENCE

An IP strategy is a subset of the business strategy (Barrett, 2002) that can be used to apply business intelligence for decision making. IP strategy plays an essential role in defining, creating, and sustaining a winning business strategy enabling value creation and strengthening multiple aspects of an effective IP strategy (Pargaonkar, 2016). In the current knowledge economy, intangible assets have gained more valuation, and hence a significant portion of enterprise value is presently governed by IP rights (Fisher & Oberholzer-Gee 2013). These IP rights, when governed wisely, yield value, and put a firm in a competitively advantageous position. The IP creation, its possession, and utilization can bring practical, long-term, and direct economic interest to nationals (Guo & Li-Hua 2008). IP strategies thus play an essential role in governing a firm's IP and are mainly aligned with the overall business strategy to successfully survive and thrive in the market place. IP rights are used to create income, to defend the firm's competitive status, and to address competitiveness (Davoudi et al., 2018). IP is a valuable financial and strategic resource that needs careful management by every organization. Without proper IP management, organizations may expose themselves to unnecessary risks and infringements as they may be unaware of the value and benefits of the IP they possess (Spruson & Ferguson, 2007).

IP strategies refer to planning related to intangible assets. Its management involves the formulation and execution of plans related to IP strategies. An appropriate IP strategy and its management enable smooth technology and knowledge transfer (Guo & Li-Hua 2008). In general, an IP management strategy includes:

1. Creating or acquiring intellectual property
2. Governing the owned intellectual property, and
3. Extracting value from the owned intellectual property

Amongst various IP rights, trade secrets, copyrights and patents can be used for protection in the ICT domain, especially for the software; however, patents are the preferred choice of firms as they provide stronger protection for the functionality of a product, the process of service (Shaikh & Londhe, 2016). Patent filing strategies can be to secure, enforce, exploit, or block, which depends on the level of innovativeness of the inventions (Süzeroğlu-Melchiors et al., 2017). Hence, patenting decisions are seen as important strategic considerations. Firms can gain maximum value from a patent depending on their ability to enforce the patent (Arrow, 1962; Holt et al. 2015; Dornelles, 2016). To enforce patents, firms need to prepare well in advance and create strategies to embed their business strategies with patenting strategies to gain a maximum advantage in the long run. Patent strategies encompass a set of resource allocation decisions and underlying "logic" of decision making about patents (Somaya, 2012). Firms seek patents to prevent copying, fence and build thickets, attaining licensing income, preventing hold-ups and rewarding R&D personnel, in addition to highlighting the innovativeness and competences of the firm (Cohen et al., 2000; Rudy & Black, 2018; Useche, 2014). Firms with active and systematic patent management outperform those that remain inactive and non-strategic (Soranzo et al., 2017)

Protection of IP does not happen automatically and may require active measures to enforce IP rights and at the same time, defend and preserve those (Spruson & Ferguson, 2007). Patent filing strategies can be used to secure, enforce, exploit, or block competition, depending on the level of innovativeness of the inventions (Süzeroğlu-Melchiors et al., 2017). Firms that remain inactive and non-strategic for patent management are outperformed by firms that have an active and systematic patent management system in place (Soranzo et al., 2017). The survival of the firms is based on how they perceive IP and patents, in particular, generate it and then utilize it further. It has become essential for firms to exploit their technologies internally as well as externally to avoid losing their value to competitors (Chesbrough, 2003). Firms can gain maximum benefit from a patent by their ability to enforce the patent (Arrow, 1962; Holt et al. 2015; Dornelles, 2016). Patent strategies include all decisions involving resource allocation along

with the logic of decision making about patents (Somaya, 2012). Firms also need to ensure that the IP they perceive and generate is aligned with their business needs and strategies to achieve long term objectives. A valid IP management strategy assists firms in capturing and protecting the outcomes of their investment in innovation. Management of intellectual property involves:

1. An understanding of what intellectual property is,
2. When the intellectual property has been created,
3. The value of the created knowledge,
4. And how to protect intellectual property that has value.

Competitive advantage over rivals is achieved by firms depending on how well they align their IP strategies with business strategies. This paper highlights the various strategies used by firms for protecting and managing their IP as available in the literature of the work carried out by researchers. It also brings forth enablers, which may be the outcome of the strategies implemented by ICT firms along with indicators of organizational performance.

2.1 Intellectual Property Management Strategies

Motohashi, (2008) defines a firm's IP strategy as "strategic use of its technology pool, which is a firm's capacity for innovation output, such as new products or processes, based on in-house R&D or acquired technology from external sources." The core purpose of an IP strategy is to develop an IP economy (Guo & Li-Hua, 2008). Without appropriate strategies, firms that are not patenting will be unable to capitalize on their investments, and researchers may be prevented from conducting even the most basic research (Clarkson & Dekorte, 2006). Hence, the role of patent management has changed from creating a purely legal barrier for competitors to a sophisticated utilization of patents to achieve maximum returns on innovation (Süzeroğlu-Melchiors et al., 2017).

IP management is the use of systematic processes to understand the intellectual property of others and to generate your own (Spruson & Ferguson, 2007). IP management strategy needs to address organizations' needs to achieve commercial goals successfully. The firms may use IP as a tool to:

- Block competing products
- Generate income from commercialization
- Deter potential infringers
- Defend an infringement action
- Attract investment
- Raise the organization's profile, or
- Increase the sale price of the organization's shares or business

IP management strategies can be viewed as offensive or defensive, depending on where and how they are applied (Spruson & Ferguson, 2007; Fisher & Oberholzer-Gee, 2013). An offensive IP strategy is generally to take action against an infringing party, while a defensive strategy is intended to obtain IP to minimize the risk of being sued by others for infringement. Striking the correct balance between being offensive and defensive is a complex task. It may depend on the market place, market size, number of players, and the technology in question. New entrants in the markets, as well as old players, can exercise both these strategies. Different strategies are listed under these two main categories are highlighted below.

2.1.1 Defensive IP Strategy

Defensive strategies seek to provide a firm the freedom to operate and commercialize its invention without hindrance from patents that belong to others (Rudy, & Black, 2018; Somaya, 2012). They are helpful when there is high fragmentation in the market for patentees, and firms are unable to arrange licensing due to transaction costs (Jell et al., 2017). Defensive strategies are thought to be reactionary, focused on protecting the current value of IP (Somaya, 2003; Rudy & Black, 2018). Various defensive IP management strategies, as highlighted below, are implemented by business firms for enhancing their organization's performance.

a) ***Legal Privilege:*** Legal privilege can be asserted by firms that do not own IP in a technology (Rudy & Black, 2018). Firms attempt to affect their competitors' patent holdings by using opposition and re-examination proceedings (Somaya, 2012). They can use legal suits to either defend the legality of the use of a technology or altogether challenge the validity of the patent holder's claim on the technology.

However, defensive litigation is a rare option as there is a high cost of litigation, along with an emotional toll. Even if a firm wins, other competitors in the market are also free to capitalize on the success, and if litigation is lost, damage awards can be huge (Fisher & Oberholzer-Gee, 2013).

b) ***Invent Around:*** Firms mainly chose to commercialize their IP possessions using in-house development and supply of goods or services based on "inventing around" a said technology. Inventing around a said technology provides an alternate way to tackle technology blockage (Cohen et al., 2000; Fisher & Oberholzer-Gee, 2013). It helps firms to increase their R&D capabilities, forms a basis for the investment in new products, a defense against others' business strategies, and a competitive advantage in the market place (Lang, 2001). However, it requires huge investments, manpower, and resources. The time taken to bring a product into the market is also longer.

c) ***Collaboration:*** Instead of inventing around solely, firms can share R&D resources by collaborating with other firms via universities, intra, and inter-industry partners who are seeking an alternative, complementing technology for the technology in question. Collaboration helps firms benefit from external knowledge partners, which facilitates the blending of external and internal ideas into new products, processes, and systems (Belderbos et al., 2014). It also helps reduce the financial burden and also distributes the risk in case of failures (Fisher & Oberholzer-Gee, 2013; Holgersson, 2012). Firms also collaborate with competitors to infiltrate their intellectual knowledge and learn about their technological skill sets (Krig & Sandra, 2017). Firms work with government and foundations in bringing out new manuals and standards in technological development. Through such collaboration, firms may emerge as leaders in technology, which maintains those standards (Krig & Sandra, 2017). Blocking patents are also common in the context of standard-setting, because once a standard is picked, any patents necessary to comply with that standard become truly essential and each patent can confer significant market power on its owner, and the standard itself is subject to holdups if these patent holders are not somehow obligated to license their patents on reasonable terms (Shapiro, 2000). Firms also collaborate to form alliances within the industry. Collaboration is built for transferring, bifurcating, or reducing the consequences of potential risk via failure in R&D output. Collaboration may also be formed in cases when there are fewer resources available for delivering technology. Collaboration efforts trigger opportunities for value creation and at the same time, also present substantial challenges in seeking to appropriate this value (Belderbos et al., 2014).

d) ***License-In:*** Licensing-in comprises procurement of required technologies under license from an IPR owner. Licensing-in is a way to acquire products or technologies without expending the time and resources necessary to develop them independently. In some cases, licensing-in is required to gain access to technologies that are proprietary but standardized in products of interest. Licensing-in reduces the time to market and might also be used to legalize infringement. For faster entry into the market place, it is recommended to license technology from the market leaders. It helps a firm to operate freely in the market without the fear of litigation. The difference in cost between acquiring knowledge from another person and originally creating that knowledge is substantial (Lindberg, 2008). Licensing can also be sought by companies for allied services required for the functioning of their product or service. By doing so, firms concentrate on the core product development and license the other dependencies from outside. Firms also license-in technology for operational freedom even if they have developed a technology in-house in case its IP is held by others. A patent license is, in

such cases, seen as "a simple means of collecting money in exchange for agreeing not to sue" (Feldman & Lemley, 2015). Licensing-in helps firms increase their business values and profits and also avoids litigation (Krig & Sandra, 2017). Firms can also coordinate the acquisition of multiple related patents using licensing to create patent fences or thickets, which later can be used as a bargaining chip in cross-licensing negotiations (Reitzig, 2007).

2.1.2 Offensive IP Strategy

Offensive patenting, on the other hand, is mostly exercised by firms having a broad patent portfolio or those owning patents of high quality. Offensive IP management strategies are thought to be proactive, focused on protecting the future value of IP (Somaya, 2003; Rudy & Black, 2018). The various offensive IP management strategies are highlighted below.

a) Exercising Market Power: As patents authorize the creation of monopolies, firms exercise market power by ensuring that no other firm infringes on its technology. The most valuable patents are not those likely to be used by the patent holder but those likely to be infringed upon by competitors because the primary role of the patent is as a bargaining chip to buy the freedom of action (Hanel, 2006). Although a patent provides its holder a right to commercialize or license its product, firms make use of enforcement mechanisms via litigation in pursuit of profits (Nerkar et al., 2007). Generally, the value of the patent right reflects the power of the patent to contribute to the profitability of the company in some manner (Holt et al. 2015). Firms employ patent litigation to detect imitation and aggressively enforce their patent's rights against possible infringement (Somaya, 2012; Rudy & Black, 2018). The use or threatened use of litigation helps a firm to protect its IP and at the same time gain competitive advantage (Rudy & Black, 2018) by enforcements with a desire to take out competition, encourage infringers to stop using patented inventions, pay higher royalties, or to build a fierce reputation (Somaya, 2012). Firms also make use of external attorneys to file patents while following a "maximization approach," resulting in more claims, filing in more countries, and more PCT applications (Süzeroğlu-Melchiors et al., 2017). Exercising market powers through litigation is high in the software industry compared to other sectors. Patent litigation is undertaken by patent holders to both dissuade and economically punish the patent infringer (Reitzig, 2007). However, patent infringement is often challenging to detect, and enforcing a patent through litigation can be extremely costly, disruptive, time-consuming, and unpredictable (Somaya, 2012).

b) Sell: Instead of capitalizing on the value of innovation, firms may also need to make trade-offs in their patent strategies to allow their technologies to create greater value in the marketplace and out compete other innovative solutions (Somaya, 2012). An outright sale is another option that can be exercised by the industry if the value of the technology is high in the hands of others (Krig & Sandra, 2017). This enables an increase in competition. Inventors can transfer their technologies to other firms within the same industry that are better suited to make the application, production, and marketing investments that are necessary to turn inventions into commercially successful innovations, by enabling combinations of resources of different types (Holgersson, 2012). Selling can also be an attractive strategy for firms if the innovator firms lack manufacturing or marketing facilities (Fisher & Oberholzer-Gee, 2013).

c) *License Out:* Licensing-out requires that the owner of IP, licenses its IP to a licensee in return for royalties and/or other considerations. It allows maximizing license revenue, thereby fully exploiting a firm's R&D capabilities (Parr & Smith, 2016). Many software vendors prefer to license the use of their product rather than sell

them, thereby retaining ownership. Licensing-out is also an enabler to ensure that the competitive firm becomes dependent on a firm's technology and does not invest in its R&D, thereby locking out the option of inventing around by competitive firms and impeding innovation (Reitzig, 2007; Krig and Sandra, 2017; Fisher & Oberholzer-Gee, 2013). Licensing-out also helps reduce the transaction costs and at the same time, may also certify invention quality to potential technology partners, thus encouraging them to license the patented technology (Somaya, 2012). Most of the time, firms patent technology with a motive to improve its bargaining position in patent licensing (Mihm et al., 2015).

d) ***Cross Licensing:*** Cross licensing is another form of barter of technology which may be royalty-free, or with a flow of royalties (Hanel, 2006). Cross licensing occurs when two competing firms with different R&D strengths take advantage of each other's intellectual assets. Cross licensing creates the same sort of synergy as a joint venture without the inconvenience and delay of setting up joint operations. These are relatively common in high technology and knowledge-led fields. Cross licensing can be a remedy to cut through patent thickets. If two patent holders are the only companies capable of manufacturing products that utilize their intellectual property rights, a royalty-free cross-license is ideal (Shapiro, 2000). Cross licensing is the preferred means by which large companies clear blocking patent positions amongst themselves or settle outstanding patent disputes (Shapiro, 2000). It is also seen as an alternative strategy for building large patent portfolios that helps to ward off patent infringement and gain access to rivals' technology (Motohashi, 2008; Fisher & Oberholzer-Gee, 2013; Rudy & Black, 2018). Patents can also be used to negotiate a cross-licensing agreement that helps in reducing the cost of acquiring the needed technology (Lang, 2001; Cockburn & MacGarvie, 2011).

e) ***Donate:*** Technology in the hands of a few helps personal gains, but when it is in the public domain it helps society. Citing this example, software companies like IBM, Google and Redhat try to donate some of their patents in the public domain (Wen et al., 2015). However, this is often done to understand how technology can be used and led further or is perceived by others. This also opens the doors of bigger firms to identify targets to acquire or collaborate in the future. Innovators may also choose to provide their innovation freely in cases where there is low return from licensing of patents due to weak protection or involving high transactional costs (Harhoff et al., 2003). It can also be disclosed freely to increase one's reputation in the market place. Donations can also act as signals of a firm's R&D capabilities, which in turn may attract financial capitals (Fisher & Oberholzer-Gee, 2013).

f) ***Signaling and Disclosure:*** Signalling technological advancements or disclosure of technology in the public domain sends signals to competitors about a firm's commitment towards a technology. This influences rivals to exit R&D competition and redirect their R&D efforts (Gill, 2008; Somaya, 2012). This may also be done by firms to generate prior art, so rival innovative firms may find it harder to obtain patents in the same technology domain, and the focal firm may be able to catch up with competitors in the race to own critical patents (Baker & Mezzetti, 2005; Somaya, 2012; Reed & Storrud-Barnes, 2011). Firms may patent "bad" inventions to mislead rivals in their efforts to build on the technologies disclosed in patents (Somaya, 2012). Specific patent actions may also be undertaken to signal the firm's patent strategy and intentions credibly. Signaling and disclosure can be done through article publication (Holgersson, 2012) using a companies' official website or web-based online publication portals such as *IP.com* or *Research Disclosure*. It is an efficient, effective, and inexpensive strategy to prevent competitors from patenting in

the technological space described in the publication disclosure (Barrett, 2002).

g) ***Patent Fencing:*** Individual patents are often ineffective as others can build technology around them (Jell et al., 2017). Firms, therefore, file patents with the sole aim of blocking competitors, ensuring freedom to operate (Hanel, 2006; Guellec et al., 2012; Weatherall & Webster, 2014). Firms try to patent not only the technology but also all related technologies of said technology, thus creating large patent portfolios (Shapiro, 2000; Lang, 2001; Weatherall & Webster, 2014; Rudy & Black, 2018). Known as "patent fencing", "patent pools", "patent stacking", "blocking", "clustering and bracketing", "blitzkrieg, consolidation", "blanketing and flooding", "fencing and surrounding", "patent harvesting and ramping up", "portfolio and network arrangements" (Jackson, 2007) or "patent thickets", the combination of multiple patents makes it costlier to invent around, and they block competitors thereby forcing competitors to license and pay higher royalties (Cohen et al., 2000; Jell et al., 2017). These patent pools help firms when threatened (or sued) over another firm's patents, as the focal firm can threaten back with its patents, leading to a situation of mutual holdup that forces a faster resolution of the standoff (Somaya, 2003; Ziedonis, 2004). Firms also use the "block to fence" strategy by acquiring a substantial number of patents not only for their core innovations but also for related processes and substitute products, hoping to drive up the cost of "inventing around" (Fisher & Oberholzer-Gee, 2013). Studies have also pointed out that the broader a firm's patent portfolio, the more likely it is to develop new products (Rudy & Black, 2018). This private strategic value of patents may be increased in the presence of 'thickets' which can help in the growth of R&D activities by constraining the ability of firms to operate without extensive licensing of complementary technologies (Noel & Schankermann, 2013) and outsiders may consider that a company with additional patents in their portfolio will have a higher future performance than a company without patents. Patent fencing is an expensive but powerful strategy to discourage or stop competitors as this tool makes it difficult for a competitor to expand on their patent portfolio without infringing on patents held by this strategy implementer (Jackson, 2007).

h) ***IP insurance:*** The need to address IP issues increases with the success of organizations as such organizations are increasingly monitored by competitors for possible infringements (Spruson & Ferguson, 2007). Business needs to protect its IP risks in-house via a legal compliance program and also by outside means via insurance. Apart from traditional insurance policies to manage risk, firms should effectively use other risk management devices, such as legal compliance programs, to ensure freedom to operate, new types of litigation insurance, and net loss insurance (Simensky & Small, 2000). Legal compliance can be used by firms to avoid infringement of others' IP and at the same time to protect their IP from infringement by others to maximize their value. However, legal compliance is rarely used in offensive or defensive roles. The cost of IP enforcement in the software domain is too expensive, and hence it is suitable for firms to insure against the financial costs of enforcement proceedings considering the significant amount of time, effort, and resources spent in creating and protecting the IP. Depending on the type of insurance and its cover, the IP insurance may cover the costs of bringing legal action to prevent or stop IP infringement by unauthorized users along with costs of legal expenses to enforce the IP right and costs of defending cross-claims brought by the alleged infringer. It may also cover the costs of proceedings brought against an organization for infringement of IP owned by a third party, including damages payable by the organization. IP insurance is advisable to firms in the early stages of IP creation, and it helps the firms to spread the risks and financial costs involved in IP lawsuits and at the same

time, acts as a deterrent to potential infringers (Spruson & Ferguson, 2007).

An offensive IP strategy is generally to exercise market power and take action against an infringing party, while a defensive strategy is intended to obtain IP to operate freely in the markets and minimize the risk of being sued by others for infringement. Having a correct balance between offensive and defensive strategies is a complex problem as it is dependent on the market place, market size, number of players, and the technology in question.

Industries are more inclined to undertake offensive or defensive strategies to enjoy positive performance outcomes (Somaya, 2003; Ziedonis, 2004; Rudy & Black, 2018). The patent strategy of firms is usually tied with its business strategies depending on its market place, market size, players involved along with the technology, and its protection. While the average patent may be a weak and porous instrument, carefully crafted patents and combinations of patents may become more effective tools for a firm's strategy (Somaya, 2012). Firms' IP strategies are evolving, and licensing decisions may be due to patent infringement, or a firm involved in a patent infringement case may adopt a serious view of IP management (Motohashi, K. 2008).

Generalizing, it can be concluded that initially, when firms do not have patents or are new entrants in a technological market, they should use a defensive approach and follow generic patenting strategies while trying to accumulate a patent portfolio. When a significant patent portfolio is available in hand, firms should try to use a more proactive offensive approach with strategic patent management that could lead to a competitive advantage (Figure 1). IP management strategy thus leads to an increase in a firm's value and its performance.

This study has several significant implications not only for IT firms but also for academics and practitioners involved in IPR, specifically in R&D and patenting. An IP strategy is driving businesses to align their business strategy with IP strategy to survive and thrive in the market place and set future goals along with competitive advantage. The present research explores various offensive and defensive IP management strategies IT firms are deploying to gain a competitive advantage in the market place. These highlighted strategies may provide the managers with an insight into various options they may deploy within their organizations to achieve a competitive advantage.

3. CONCLUSION

IP in the field of ICT is gaining importance with the advent of new emerging technologies. Creating and managing IP in the field of ICT has become a key differentiator for the success of ICT firms as the industry is moving with a rapid pace of innovations that have a shorter life cycle. The exploitation of IP and patents in particular is often linked with business sales, export quality, and marketing needs, along with research direction strategies to ensure that a firm remains competitive in business.

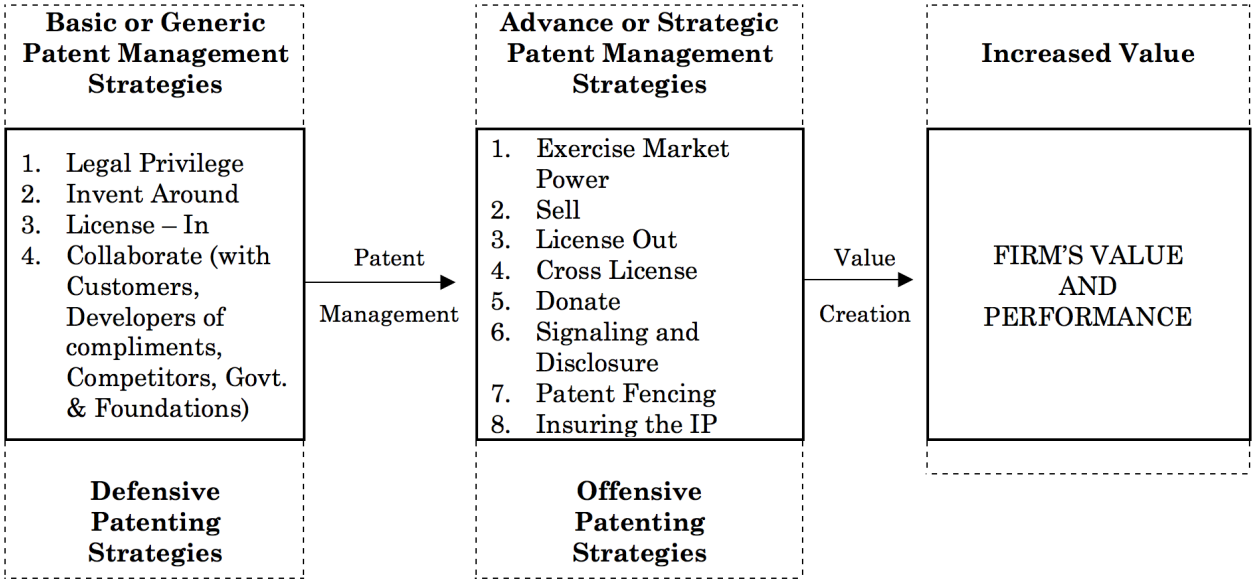


Figure 1 Patenting strategies and firm's value.

Firms have started looking and opting for various IP management strategies to achieve success and competitive advantage. IP strategy has thus become a force for organizational performance, and businesses have begun aligning their IP strategy with their business strategy to successfully survive and thrive in the market place. Amongst various IP rights, trade secrets, copyrights, and patents can be used for protection in the ICT domain, especially for software; however, patents are the preferred choice of firms as they provide stronger protection for the functionality of a product, the process of service. It is also seen that firms with active and systematic patent management outperform those that remain inactive and non-strategic.

Various offensive and defensive IP strategies exist with the aim of attaining a competitive edge in the market place. Defensive strategies seek to provide a firm the freedom to operate and commercialize an invention without hindrance from patents that belong to others. Defensive strategies are thought to be reactionary, focused on protecting the current value of IP. Offensive patenting, on the other hand, is mostly exercised by firms having large patent portfolios or those owning patents of high quality. Offensive IP management strategies are thought to be proactive, focused on protecting the future value of IP. Industries are more inclined to undertake an offensive or defensive strategy to enjoy positive performance outcomes.

IPR in general and patents in particular serve as a barter system that helps promote innovation and research by putting innovation in the public domain in exchange for exclusive rights over the said technology for a limited period. The creation, protection, and enforcement of IP can bring direct, practical long-term economic interest to nations. Firms seek to gain and maintain a competitive advantage by managing and protecting IP as they accumulate patent portfolios to gain market share or increase profits via multiple strategies. The study puts forth various IP strategies used by firms. Many of these strategies are still evolving and are implemented proactively or reactively depending on various scenarios and situations.

4. REFERENCES

Arrow, K.J. (1962). Economic welfare and the allocation of resources for invention. The Rate and Direction of Inventive Activity: Economic and Social Factors. Princeton Univ. Press, Princeton.

Baker, S., & Mezzetti, C. (2005). Disclosure as a Strategy in the Patent Race. Journal of Law and Economics, 48(1), 173-194.

Barrett, B. (2002). Defensive use of publications in an intellectual property strategy. Retrieved from http://biotech.nature.com.

Belderbos, R., Cassiman, B., Faems, D., Leten, B., & Van Looy, B. (2014). Co-ownership of intellectual property: Exploring the value-appropriation and value-creation implications of co-patenting with different partners. Research Policy, 43(5), 841-852.

Cameron, D. M., MacKendrick, R. S., & Chumak, Y. (2006). Patents for Computer Implemented Inventions and Business Methods. Canadian IT Law Association; Toronto. Retrieved from http://www.jurisdiction.com/itpatents.pdf

Chesbrough, H.W. (2003) Open Innovation: The New Imperative for Creating and Profiting from Technology, Harvard Business School Press, Boston.

Clarkson G & Dekorte D. (2006). The problem of patent thickets in convergent technologies, New York Academy of Sciences, 1093, 180–200.

Cockburn, I., M., & MacGarvie, M. J. (2011). Entry and Patenting in the Software Industry. Management Science, 57(5), 915-933.

Cohen, W.M., Nelson, R.R., and Walsh, J.P. (2000). Protecting their intellectual assets: appropriability conditions and why us manufacturing firms patent (or not), (No. w7552). National Bureau of Economic Research, Cambridge, MA.

Davoudi, S. M. M., Fartash, K., Zakirova, V. G., Belyalova, A. M., Kurbanov, R. A., Boiarchuk, A. V., & Sizova, Z. M. (2018). Testing the Mediating Role of Open Innovation on the Relationship between Intellectual Property Rights and Organizational Performance: A Case of Science and Technology Park. Eurasia Journal of Mathematics, Science and Technology Education, 14(4), 1359-1369.

Debackere, K., and Veugelers, R. (2005). The role of academic technology transfer organizations in improving industry science links, Research Policy, 34(3), 321–342.

Dedrick J. & Kraemer, K. L. (1993). Information Technology in India: The quest for self-reliance. Asian Survey, 33(5), 463-492.

Dornelles, J. (2016). Why are they hiding? Patent secrecy and patenting strategies. Retrieved from http://dee.uib.es/digitalAssets/410/410898_jmp---juliana-pavan-dornelles.pdf

EPO (2013), Patents for software? European law and practice, Retrieved October 11, 2013, from European Patent Office, Web site: http://www.epo.org

Feldman, R., & Lemley, M. A. (2015). Do patent licensing demands mean innovation. Iowa Law Review, 101(1), 137.

Fisher III, W. W., & Oberholzer-Gee, F. (2013). Strategic management of intellectual property: an integrated approach. California management review, 55(4), 157-183.

Gill, D. (2008). Strategic disclosure of intermediate research results. Journal of Economics & Management Strategy, 17(3), 733-758.

Guellec, D., Martinez, C., & Zuniga, P. (2012). Pre-emptive patenting: securing market exclusion and freedom of operation. Economics of Innovation and New Technology, 21(1), 1-29.

Guo, M., & Li-Hua, R. (2008). Conceptual framework of strategic intellectual property management: a case study of Henan Province, China. Journal of Technology Management in China, 3(3), 307-321.

Hanel, P. (2006). Intellectual property rights business management practices: a survey of the literature, Technovation, 26(8), 895-931.

Harhoff, D., Henkel, J., & Von Hippel, E. (2003). Profiting from voluntary information spillovers: how users benefit by freely revealing their innovations. Research Policy, 32(10), 1753-1769.

Holgersson, M. (2012). Innovation and Intellectual Property: Strategic IP Management and Economics of Technology. Ph.D. thesis, Chalmers University of Technology, Gothenburg, Sweden.

Holt, K. F., O'Shaughnessy, B. P., & Herman, T. B. (2015). What's It Worth: Principles of Patent Valuation. Landslide, 8, 33.

IPO (2013), Guidelines for Examination of Computer Related Inventions (CRIs); Office of the Controller General of Patents, Designs and Trademarks ; 3. India.

Isa, D., Blanchfield, P., & Chen, Z. (2009). Intellectual Property Management System for the Super-Capacitor Pilot Plant. In IC-AI. 708-714.

Jackson, PJ (2007). The dangers of patents as weapons. LLM thesis. University of Kent, Canterbury, UK.

Jell, F., Henkel, J., & Wallin, M. W. (2017). Offensive patent portfolio races. Long Range Planning, 50(5), 531-549.

Jyoti, Banwet, D.K., & Deshmukh, S.G. (2010). Modelling the success factors for national R&D organizations: a case of India. Journal of Modelling in Management, 5(2), 158-175.

Kavida, V., & Sivakoumar, N. (2008). Intellectual property rights-the new wealth of knowledge economy: An Indian perspective. SSRN 1159080. Retrieved December 12, 2015, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1159080

Khode, A., & Jambhorkar, S. (2017). A Literature Review on Patent Information Retrieval Techniques. Indian Journal of Science and Technology, 10(36), 1-13.

Krig, M. L., & Sandra, L. (2017). Business Value Enhancing Factors of Aligning IP Strategy with Corporate Strategy. Master Thesis, Blekinge Institute of Technology, School of Management, Sweden.

Lang, C. J. (2001). Management of intellectual property rights: Strategic patenting. Journal of Intellectual Capital, 2(1), 8-26.

Lindberg, V. (2008). Intellectual property and open source: a practical guide to protecting code. O'Reilly Media, Inc.

Mahajan V., Nauriyal D.K., Singh S P. (2015). Trade performance and revealed comparative advantage of Indian pharmaceutical industry in new IPR regime, International Journal of Pharmaceutical and Healthcare Marketing, 9(1), 56-73.

McGowan, M. K., Stephens, P., & Gruber, D. (2007). An exploration of the ideologies of software intellectual property: The impact on ethical decision making. Journal of Business Ethics, 73(4), 409-424.

Mihm, J., Sting, F. J., & Wang, T. (2015). On the effectiveness of patenting strategies in innovation races. Management Science, 61(11), 2662-2684.

Motohashi, K. (2008). Licensing or not licensing? An empirical analysis of the strategic use of patents by Japanese firms. Research Policy, 37(9), 1548-1555.

Nerkar, A., Paruchuri, S., & Khaire, M. (2007). Business method patents as real options: Value and disclosure as drivers of litigation. Real Options Theory, 24, 247-274.

Noel, M., & Schankerman, M. (2013). Strategic patenting and software innovation. The Journal of Industrial Economics, 61(3), 481-520.

OECD (2008), report number DAF/COMP(2007)40. Accessed on October 176, 2019 from https://www.oecd.org/competition/abuse/39888509.pdf

Pargaonkar, Y. R. (2016). Leveraging patent landscape analysis and IP competitive intelligence for competitive advantage. World Patent Information, 100(45), 10-20.

Parr, R. L., & Smith, G. V. (2016). Intellectual Property: Valuation, Exploitation, and Infringement Damages, 2016 Cumulative Supplement. John Wiley & Sons.

Porter, M. E. (1993). The competitive advantage of nations, 73-93. Harvard Business School Management Programs. Cambridge.

Rao S. S. (2001). IPR in the ensuing global digital economy, Library Hi-Tech, 19(2), 179-185.

Reed, R., & Storrud-Barnes, S. F. (2011). Patenting as a competitive tactic in multipoint competition. Journal of Strategy and Management, 4(4), 365-383.

Reitzig, M. (2007). How executives can enhance IP strategy and performance, MIT Sloan Management Review, 49(1), 37-43.

Rudy, B. C., & Black, S. L. (2018). Attack or defend? The role of institutional context on patent litigation strategies. Journal of Management, 44(3), 1226-1249.

Shaikh S. A., Londhe B. R. (2016), Intricacies of Software Protection: A Techno - Legal Review, Journal of Intellectual Property Rights, 21, 157-165.

Shaikh, S.A. and Singhal, T.K. (2018) Business intelligence through patent filings: An analysis of IP management strategies of ICT companies. Journal of Intelligence Studies in Business. 8(2) 62-76.

Shapiro, C. (2000). Navigating the patent thicket: Cross licenses, patent pools, and standard-setting. Innovation policy and the economy, 1, 119-150.

Simensky, M., & Small, L. A. (2000). The Management of Intellectual Property Risks. Handbook of Business Strategy, 1(1), 125-138.

Somaya, D. (2003), "Strategic determinants of decisions not to settle patent litigations," Strategic Management Journal, 24(1), 17-38.

Somaya, D. (2012). Patent strategy and management: An integrative review and research agenda. Journal of Management, 38(4), 1084-1114.

Soranzo B., Nosella A., Filippini R, (2017). Redesigning patent management process: an Action Research study, Management Decision, 55(6), 1100-1121.

Spruson, D., & Ferguson, N. (2007). Intellectual property management: A practical guide for electrical and Electronics related industries. Australian Government.

Süzeroğlu-Melchiors, S., Gassmann, O., & Palmié, M. (2017). Friend or foe? The effects of patent attorney use on filing strategy vis-a-vis the effects of firm experience. Management Decision, 55(6), 1122-1142.

Torvinen, P., & Väätänen, J. (2014). External technology commercialisation and markets for technology in Russian manufacturing industry. International Journal of Technology Marketing, 10(1), 4-24.

The United States. Federal Trade Commission. (2003). To promote innovation: The proper balance of competition and patent law and policy. DIANE Publishing. Accessed on October 176, 2019 from https://www.ftc.gov/sites/default/files/documents/reports/promote-innovation-proper-balance-competition-and-patent-law-and-policy/innovationrpt.pdf

Useche, D. (2014). Are patents signals for the IPO market? An EU–US comparison for the software industry. Research Policy, 43(8), 1299-1311.

Wang H, Mingyong Lai, Maxim Spivakovsky, (2012), Does IPR promote innovation? New evidence from developed and developing countries, Journal of Chinese Entrepreneurship, 4(2), 117-131.

Weatherall, K., & Webster, E. (2014). Patent enforcement: a review of the literature. Journal of Economic Surveys, 28(2), 312-343.

Wen, W., Ceccagnoli, M., & Forman, C. (2015). Opening up intellectual property strategy: Implications for open source software entry by start-up firms. Management Science, 62(9), 2668-2691.

WIPO-b (World Intellectual Property Organisation) Intellectual Property Handbook; Chapter 2 - Fields of Intellectual Property Protection, Page 17, Section 2.1 and 2.5

Zahra, S.A.; Nielsen, A.P. (2002) Sources of capabilities, integration and technology commercialization, Strategic Management Journal, 23(5), 377–398.

Zhang, H., & Yang, X. (2016). Intellectual property rights protection and export quality. International Journal of Development Issues, 15(2), 168-180.

Ziedonis, R. H. (2004). Don't fence me in: Fragmented markets for technology and the patent acquisition strategies of firms. Management Science, 50(6), 804-820.

Journal of Intelligence Studies in Business
Vol. 9, No. 2 (2019) pp. 43-58
Open Access: Freely available at: https://ojs.hh.se/

Business intelligence using the fuzzy-Kano model

Soumaya Lamrhari[a]*, Hamid Elghazi[b] and Abdellatif El Faker[a]

[a]*ENSIAS, Mohammed V University Rabat, Morocco*
[b]*National Institute of Posts and Telecommunications Rabat, Morocco*

Corresponding author (*): soumaya_lamrhari@um5.ac.ma

Received 25 September 2019 Accepted 24 October 2019

ABSTRACT Today, understanding customer satisfaction is becoming a difficult and complex task for companies due to the explosive growth of the voice of the customer in online reviews. This has pushed companies to rethink their business strategies and resort to business intelligence techniques in order to help them in analyzing customer requirements and market trends. This paper proposes a decision support framework for dynamically transforming the voice of the customer data into actionable insight. The framework measures the customer satisfaction by extracting key products' aspects along with customers' sentiments from online reviews using a text mining technique: the latent Dirichlet allocation approach. We apply the Fuzzy-Kano model to classify the real customer requirements, then, map them dynamically to the SWOT matrix. The proposed approach is extensively tested on an empirical dataset based on several performance metrics including accuracy, precision, recall, and F-score. The reported results showed that latent Dirichlet allocation approach has correctly extracted aspects with 97.4% accuracy and 92.4 % precision.

KEYWORDS Business intelligence, customer satisfaction, decision support framework, Fuzzy-Kano model, latent Dirichlet allocation, online reviews, text mining, voice of the customer, web intelligence

"The secret of successful retailing is to give your customers what they want."

Sam Walton

1. INTRODUCTION

In today's competitive marketplace, business leaders have realized that customers are the major driving force leading a company to thrive (Carulli et al., 2013) (Lee et al., 2014). In fact, most of the product-based companies require an in-depth understanding of their customers' satisfaction. Thus, they resort to business intelligence (BI) techniques in order to provide competitive products that meet the customer needs and go in line with the current market trend (Sabanovic and Søilen, 2012). The voice of the customer (VOC) is a widely used term in market research that describes the customers' feedback about their expectations and experiences in relation to products and services. This is considered an essential first step in developing a successful product or service (Aguwa et al., 2012). The VOC is usually captured in a variety of ways such as questionnaire surveys, face to face interviews, telephone interviews, and discussion groups (Goodman, 2014) (Rese et al., 2015). However, most of these methods are demanding in terms of time, cost, and their geographic reach (Szolnoki and Hoffmann, 2013). Additionally, the participants' willingness to provide actual input can impact the collected data quality (Reyes, 2016). Besides, the surveys are generally conducted occasionally, which makes

the timeliness of the gathered data questionable (Culotta and Cutler, 2016). Consequently, we need to consider other alternative data sources to reveal customer expectations.

The growing popularity of social media and BI in the last decade makes them a valuable digital channel for listening and capturing customers' voices (Gioti et al., 2018). Unlike conventional approaches, the VOC on social media is publicly available, easily accessible anywhere and anytime at low cost. Examples of these VOCs include customer posts, comments, and reviews. Customer reviews can be considered a trustworthy VOC since they hold massive data where customers voluntarily share their experiences about a specific product or service after use or purchase. Unfortunately, these reviews may not explicitly reflect customer needs since they require more advanced data analysis methods. Therefore, most companies have adopted BI techniques (Nyblom et al., 2012), such as text mining, to discover hidden patterns in this large amount of textual data to support the decision making process (Søilen et al., 2017) (Xu and Li, 2016) (Jia, 2018).

Plenty of studies have been conducted to explicitly or implicitly understand customer satisfaction from online review content. For instance, Decker and Trusov (2010) applied an econometric framework based on Poisson regression, binomial regression, and latent class Poisson regression models. The basic potential of using those classification algorithms is to estimate the relative strength of effects resulting from the list of attributes identified through customer reviews about mobile phones. The methodology findings reveal that the negative binomial regression approach provides significant estimation parameters, which quantify the effects that the product attributes have on overall customer satisfaction. Park and Lee (2011) proposed a systematic framework for extracting customer requirements from an online customer center and transforming them into product specifications data. In their approach, customer opinions are collected, then a text mining analysis is conducted on customer complaints to extract meaningful keywords. Based on the extracted VOCs, customers are clustered into different groups with similar needs. Then, the target groups will be carefully selected by the companies. Further, a co-word and a decision tree analysis are used to translate the customer requirements into product specifications. Xiao et al. (2016) established a novel econometric preference measurement model for extracting overall customers' preferences from online product reviews. The model allows a semi-automatic extraction of product features along with the related reviewers' sentiments. Then, aggregate customer preferences are extracted from online product reviews by a modified ordered choice model, which considers the variety of customers' ratings and allows them to assign rating sores with their own thresholds. Furthermore, the identified customer requirements are classified into different categories, e.g. basic, performance, excitement, innovation-needed, reverse and divergent, by using a marginal effect-based Kano model, which is an extension of the classical Kano model that employs the marginal effect information disclosed by the proposed modified ordered choice model.

In addition, other research studies have applied an aspect-based sentiment analysis approach for understanding customers' satisfaction. This approach involves extracting aspects and finding their corresponding sentiments. Latent Dirichlet allocation (LDA) is considered a state-of-the-art modeling tool for extracting products' features in the aspect-based sentiment analysis (Saura et al., 2019). For instance, Farhadloo et al. (2016) proposed a Bayesian approach that models the customer satisfaction based on the individual aspect ratings. First, the study utilizes the aspect-based sentiment analysis method described in (Farhadloo and Rolland, 2013) as a basis to transform unstructured input data into semi-structured data. Then, the Bayesian method enables the extraction of the relative importance of each aspect of the product or service. For consumer-generated content in marketing, Tirunillai and Tellis (2014) proposed a unified framework that extracts the key latent quality dimensions (known as a "topic" in the LDA literature) of consumer satisfaction and the associated sentiments using unsupervised Bayesian learning algorithm based LDA. Moreover, the approach determines the validity, importance, dynamics, and heterogeneity of the extracted dimensions. In another context, Guo et al. (2017) put forward an LDA based approach to identify the most important dimensions of customer service in the hotel sector. Then, they performed a perceptual mapping to represent the key dimensions influencing the visitors' satisfaction and the visitors' perceived ratings

in different hotel classification. Qi et al. (2016) proposed an automatic filtering model to mine customers' requirements from online reviews. First, it filters out the reviews that are helpful for product improvement. Then, a lexicon-based sentiment analysis, LDA, and page rank are used to rank the terms based on their frequencies and semantic relationships. In addition, the conjoint analysis and the Kano model are utilized to determine the product attribute weights and categories and evaluate their impact on customer satisfaction.

Despite the contributions made by the aforementioned studies regarding the understanding of customer satisfaction from online reviews, they still have some drawbacks. First, in (Decker and Trusov, 2010), (Farhadloo et al., 2016), (Qi et al., 2016), (Xiao et al., 2016); (Park and Lee, 2011), the authors quantified the effects that customer requirements may have on their satisfaction by using various modeling methods that measure product attributes, e.g. weights and importance. While in (Guo et al., 2017), (Tirunillai and Tellis, 2014), the authors focused only on mining the relevant products' attributes. Second, most of the existing studies that have measured the effects of customer requirements on customer satisfaction have not classified the identified requirements either from the customer or the provider perspectives. Third, our approach bears a close resemblance to the one proposed by Qi et al. (2016), except that in our study, we have incorporated the Fuzzy analysis to the Kano model instead of the conjoint analysis. With Fuzzy analysis, the measurement of each product's attribute is presented in the form of the degree of membership allowing the customers to express their preferences towards multi-attributes at the same time, unlike the conjoint analysis where the customers can only express their preferences for a single attribute.

Based on the results reported in (Tirunillai and Tellis, 2014), (Qi et al., 2016), (Guo et al., 2017), LDA has demonstrated good stability and satisfactory performance in terms of accurately extracting the key customer requirements from a large volume of online reviews. Therefore, we have selected it as a topic modeling method in our approach. To the best of our knowledge, this is the first attempt to combine LDA, the Fuzzy-Kano model and the SWOT method into one decision support framework for understanding customer satisfaction. Specifically, we will analyze the collected VOC from online reviews, then, extract the actual customers' requirements that have more impact on their experiences with a given product or service.

Such a framework is beneficial for companies since it allows them to deeply understand the customers' needs and proactively adapt their product/service or even their business model accordingly. It is composed of four major modules. The first one consists of collecting and preprocessing data from online customer reviews. The second one extracts the products' aspects and the corresponding customers' sentiments from the preprocessed data using LDA. The third module classifies the real customer needs that affect their satisfaction based on the Fuzzy-Kano model. The fourth module maps the Fuzzy-Kano model's output to a SWOT matrix in order to easily interpret the obtained results. The proposed approach is extensively evaluated using an empirical dataset, which includes mobile phone reviews collected from Amazon. The evaluation is based on several performance metrics including accuracy, precision, recall, and F-score.

The remainder of this paper is organized as follows. Section II provides the theoretical background of the proposed framework. Section III describes our methodology. In Section IV, we evaluate the effectiveness of our method using a real case study. In section V, we draw some conclusions and shed light on further research directions.

2. THEORETICAL BACKGROUND

2.1 Latent Dirichlet Allocation (LDA)

In this paper, we seek a way to map customers' reviews to the topics, without having prior knowledge on what those topics are. This calls into question the unsupervised classification problem on natural language. LDA is an unsupervised topic modeling approach widely applied in natural language processing. The present study deployed LDA (Blei, 2012) instead of other topic model approaches found in the literature because it relies on more comprehensive probabilistic assumptions on the text generation and has shown satisfactory performance and good stability when classifying large data sets (Lu et al., 2011) (Alghamdi and Alfalqi, 2015) (Hofmann, 2017). In LDA, each document consists of a mixture of topics and each topic consists of a collection of words. Given a corpus D consisting of M documents each of length N, each document contains a sequence of W words, each of these words represents the v^{th} word in a vocabulary

of V distinct terms and K is the total number of topics. Thus:

- α and β define the prior distribution parameters per-document topic distribution and per-topic word distribution respectively.
- θ_m is the topic distribution for document m.
- φ_k is the word distribution for topic k.
- z_{nm} is the topic for the n^{th} word in document m.
- and w_{mn} is the specific word

Formally, LDA generates a corpus D of M documents according to the following generative process:

- Choose a topic distribution $\theta_i \sim Dir(\alpha)$, where $i \in \{1, \ldots, M\}$, and $Dir(\alpha)$ is a Dirichlet distribution with scaling parameter α which typically is sparse ($\alpha < 1$).
- For each topic $k \in \{1, \ldots, K\}$, Choose $\varphi_k \sim Dir(\beta)$, where β is typically sparse.
- For each of the word positions i, j, where $j \in \{1, \ldots, N_i\}$, and $i \in \{1, \ldots, M\}$:
  - Choose a topic $z_{i,j} \sim Multinomial(\theta_i)$.
  - Choose a word $w_{i,j} \sim Multinomial(\varphi_{z_{i,j}})$.

Moreover, a graphical model can also mirror the generative process of documents. As depicted in Figure 1, the boxes refer to repeated contents where the number of repetitions is presented by the variable at the corner of the corresponding box. The blue node represents the only observed variable (w). The white nodes denote latent variables (φ, θ); Gray nodes represent hyperparameters (α and β). The arrows indicate dependencies among the model parameters.

Practically, the model must determine the hidden variables from the data, namely the document-topic distribution θ, and the topic-word distribution φ. To this end, the Gibbs Sampling algorithm (Darling, 2011) is applied to estimate those two LDA parameters.

2.2 Kano Model

The Kano model (Kano, 1984) is an effective tool used by companies to integrate the VOC into the product and service development

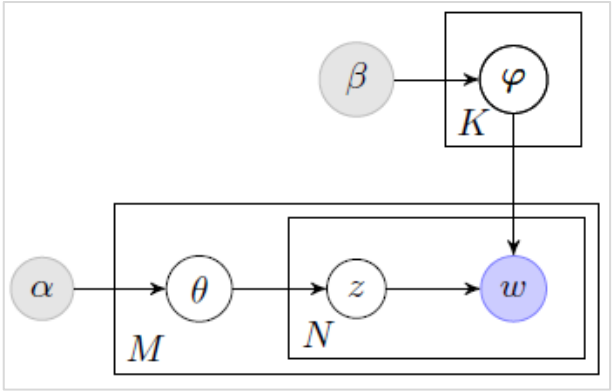


Figure 1 The graphical representation of the LDA model, redrawn from (Blei, 2012)

lifecycle. It is regarded as a nonlinear relationship between product quality and customer satisfaction. It measures customer sentiments to discover which customer requirements have the highest impact on customer satisfaction (Tontini et al., 2013).

The Kano model often carries out surveys and questionnaire investigations on customers to determine the requirements of a particular product or service. For a given product's aspect, a functional question (aspect's presence) and a dysfunctional question (aspect's absence) are asked. Each question form should be answered on a five-point scale such as: like, necessary, neutral, unnecessary, and dislike. Based on a statistical analysis of all the accumulated responses of the survey, each answer pair is aligned with the Kano evaluation (Table 1), forming certain requirements (Ullah and Tamaki, 2011). Table 1 shows that by combining the two answers (functional and dysfunctional), the product's aspects can be classified into six categories of requirement that influence customer satisfaction, including:

- "Must-be" (M) requirement is expected by the customers, its presence does not lead to customer satisfaction, but its absence leads to extreme customer dissatisfaction.

Table 1 The standard Kano evaluation (Ullah and Tamaki, 2011). Nec = necessary; Neu = neutral; Unnec = unnecessary; Dis = dislike.

		Dysfunctional				
		Like	**Nec.**	**Neu**	**Unnec**	**Dis**
Functional	**Like**	Q	A	A	A	O
	Nec	R	I	I	I	M
	Neu	R	I	I	I	M
	Unnec	R	I	I	I	M
	Dis	R	R	R	R	Q

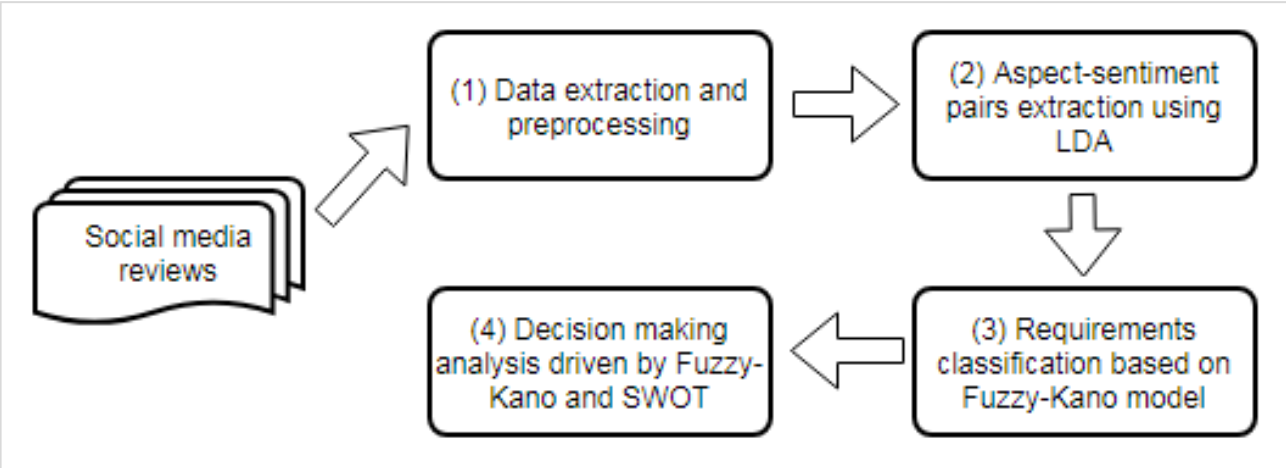


Figure 2 The proposed decision support framework.

- "One-dimensional" (O) requirement is the property of a customer need that increases customer satisfaction when it is fulfilled. Inversely, customer satisfaction decreases when it is not fulfilled.
- "Attractive" (A) requirement is usually uncommon or unexpected by the customers, if included, can truly increase customer satisfaction; if not, there is no feeling of dissatisfaction.
- "Indifferent" (I) requirements are those that the customer does not care about whether they exist or not. That is, these attributes will cause neither the satisfaction nor the dissatisfaction of customers, but that does not mean they do not impact the company's production decisions.
- "Reverse" (R) requirements are those whose presence results in dissatisfaction since not all customers are alike. In other words, what makes one customer satisfied might probably alienate another.
- And the "Questionable" (Q) requirement, which occurs when the customer selects an unclear answer from both functional and dysfunctional sides.

In addition, the Kano questionnaires and surveys allow the users to select only a single option from a set of options. That makes them unable to express their uncertainty toward certain aspects by selecting more than one choice. To address the issue of uncertainty concerning people's satisfaction as well as the vagueness of human thought, our study combines the classical Kano model with the fuzzy analysis to obtain an equivalent Fuzzy-Kano model that classifies the customers' requirements based on fuzzy logic rather than binary logic (Lee and Huang, 2009). The Fuzzy-Kano model allows customers to express multi-feeling, with the help of the different Kano categories, by giving fuzzy satisfactory values to certain aspects. This fuzzy set of values is represented by variable membership degrees ranging from 0 to 1, reflecting the uncertainty, where the sum of elements is equal to 1. Furthermore, this approach automates the building of the Kano model. It incorporates the VOCs into the Fuzzy-Kano model through LDA to obtain much larger scale data with more reliable insights since the classical Kano model, when used alone, cannot directly handle such data.

3. METHODOLOGY

The proposed framework is composed of four modules as illustrated in Figure 2: (1) data extraction and preprocessing; (2) aspect-sentiment pairs extraction using LDA; (3) requirements classification based on the Fuzzy-Kano model; and (4) decision-making analysis driven by Fuzzy-Kano and SWOT. In this section, we describe each of these modules.

3.1 Data Extraction and Preprocessing

The first module consists of gathering online customer reviews as the material for analysis and saving them in the form of a table in which each review denotes a document. Generally, reviews contain emoticons, special characters, punctuation, HTML tags, capital letters and misspelled words. So, it is necessary to apply a

set of operations to each review before moving to the next module. These preprocessing operations include:

Tokenization: is the act of breaking up a sequence of textual content into words, phrases, and symbols called tokens. These tokens are used as input data for further processing.

Stop word removal: is the process of filtering out irrelevant words and characters from data, such as prepositions and pronouns.

Part-Of-Speech Tagging (POST): is applied to assign a special label to each token (word) in a text such as a noun, verb, or adjective.

Filtering tokens: is used to filter out all words where the length is out of the range [2-25 characters].

Transforming cases: consists of converting all tokens into lowercase.

Stemming: is applied to discard affixes from each word to obtain their root form.

Additionally, some reviews can be wrapped in a specific electronic file format, such as HTML, XML or JSON, which sometimes requires transformation into another format so as to be easily processed by the next modules. After performing the aforementioned preprocessing operations, a set of valid words is generated by excluding all meaningless words from the token list. Thus, a document-term matrix is produced, which indicates terms and their occurrence frequencies in each document.

3.2 Aspect-Sentiment Pairs Extraction using LDA

In this module, we begin by implementing LDA to reveal all topics being discussed by customers in the reviews. For this, we compute the probability of each word in the review as written in equation 1:

$$p(w|R) = \sum_{i=1}^{K} p(w|T) \times p(T|R_i) \qquad (1)$$

Where $p(w|T)$ is the probability of a word w given a topic T and $p(T|R_i)$ is the probability of a topic T given a review R_i, with K is the total number of reviews in the overall collection.

Then, we extract aspects and sentiments that appear together in the same topic distribution according to the POS tagging process. Words describing sentiments are mainly represented by adjectives and adverbs, meanwhile, a product aspect is mainly represented by nouns or noun phrases (Hu and Liu, 2004a), but not all nouns refer to aspects. Therefore, we select first the most representative nouns as aspect candidates according to their co-occurrence frequencies in the review, as well as their appearance with sentiment words. To identify sentiment word orientation, the Wordnet (Miller, 1995) is used as well as the opinion lexicon provided in (Hu and Liu, 2004b), when the sentiment words are not supported by Wordnet. Next, we use the popular approach of Hu and Liu (2004b) to construct aspect-sentiment pairs, which is based on extracting nearby adjectives to a frequent aspect.

Practically, we define a nearby adjective as the nearest opinion word to a specific aspect considering token distance (measured in the number of words far away from that aspect). The maximum number of the nearest sentiment words is set at two for the simple reason that usually when a third word is found, it was certainly describing another aspect that was ignored during processing. By doing so, we prevent the incorrect attribution of a sentiment word to an aspect. Moreover, we consider that once a sentiment word is assigned to an aspect, it will not be considered in the future attribution.

To compute the final sentiment score for an aspect (positive or negative), we sum up all sentiment word scores related to that aspect as follows:

$$A_i.ss = \sum_{j} \frac{SW_j.ss}{dist(SW_j, A_i)} \qquad (2)$$

Where $A_i.ss$ is the sentiment score of an aspect A_i, $SW_j.ss$ is the polarity score $\{-1,1\}$ given to the j^{th} sentiment word according to the opinion lexicon, and $dist(SW_j, A_i)$ is the distance between the aspect A_i and the identified sentiment word SW_j. This allows us to identify the opinion words with the highest weight, i.e. the nearest opinion word to the aspect.

3.3 Requirements Classification based on Fuzzy-Kano model

In this module, we use the aspect-sentiment pairs generated previously in combination with the Fuzzy-Kano model to classify the real customer requirements that affect customer satisfaction. In the document collection, each comment is written by a customer, c, to express a sentiment, s, toward several aspects asp of an item, i. By using the quadruplet $\{s, i, asp, c\}$, we form the matrix of aspect and sentiment distribution, denoted as $A = (a_{ij})_{1\le j\le q}^{1\le i\le p}$. For instance, in equation 3, rows represent aspects and columns denote items. The matrix entries represent the customer's sentiment c_{pq} toward the aspect p of the item q. We assign +1 to a positive attitude, -1 to a negative attitude, and 0 to a neutral attitude or no opinion expressed. Then, we construct for each aspect a set of n-dimensional vector distributions. For example, the first row in the matrix indicates that for aspect 1, the customer marks a negative attitude for item 1, neutral or no feeling toward item 2, and a positive attitude for item q. Thus, each row in the matrix constitutes a customer's sentiment vector corresponding to that aspect.

$$A = \begin{bmatrix} -1 & 0 & \cdots & 1 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & 1 \end{bmatrix} \qquad (3)$$

To apply the Fuzzy-Kano, first we calculate for each aspect the customer's degree of preference when the aspect has a functional presence and the customer's degree of dislike when the aspect has a dysfunctional absence or insufficiency. Probability gives real knowledge when the customer feelings are ambiguous or uncertain. So, we calculate such degrees as probabilities of preference and dislike. They are represented, respectively, in equations 4 and 5:

$$preference(c, Asp_i) = \frac{N_s}{p \times q} \times \frac{S_i^+}{S_i} \qquad (4)$$

$$dislike(c, Asp_i) = \frac{N_s}{p \times q} \times \frac{S_i^-}{S_i} \qquad (5)$$

Where $preference(c, Asp_i)$ and $dislike(c, Asp_i)$ represent the probabilities that customer, c, has a positive or negative sentiment, respectively, for aspect Asp_i for a specific item, N_s denotes the number of sentiments either positive or negative expressed by a customer, c, toward some aspects, $p \times q$ refers to the dimension of aspect-sentiment matrix, S_i^+ and S_i^- represent the number of positive and negative sentiments given by c for aspect Asp_i respectively, and S_i is the total number of sentiment attitudes expressed by several customers for the aspect Asp_i.

Second, each of the obtained preference and dislike values refers to a fuzzy set, which contains elements that have varying degrees of membership in the set. These degrees correspond to the five Kano's standard answers ('like', 'necessary', 'neutral', 'unnecessary', and 'dislike'). They are determined using the membership functions where each element of the fuzzy set is mapped to a value ranging from 0 to 1. In particular, we employ in this paper the triangular membership function because of its simplicity in determining the input parameter values, namely the *preference* and *dislike* in our case (Umoh and Isong, 2013). According to the triangular membership method, the five Kano's standard answers are represented as five triangular fuzzy numbers between $\tilde{0}$ and $\tilde{1}$, as follows:

- Dislike: $(0, 0, 0.25)$

$$\mu_R(x) = \begin{cases} 0.25 - x & 0 \le x \le 0.25 \\ 0 & otherwise \end{cases}$$

- Unnecessary: $(0, 0.25, 0.5)$

$$\mu_R(x) = \begin{cases} x & 0 \le x \le 0.25 \\ 0.5 - x & 0 \le x \le 0.5 \\ 0 & otherwise \end{cases}$$

- Neutral: $(0.25, 0.5, 0.75)$

$$\mu_R(x) = \begin{cases} x - 0.25 & 0.25 \le x \le 0.5 \\ 0.75 - x & 0.5 \le x \le 0.75 \\ 0 & otherwise \end{cases}$$

- Necessary: $(0.5, 0.75, 1)$

$$\mu_R(x) = \begin{cases} x - 0.5 & 0.5 \le x \le 0.75 \\ 1 - x & 0.75 \le x \le 1 \\ 0 & otherwise \end{cases}$$

- Like: $(0.75, 1, 1)$

$$\mu_R(x) = \begin{cases} x - 0.75 & 0.75 \le x \le 1 \\ 0 & otherwise \end{cases}$$

Where x is the fuzzy set represented by the degree of preference/dislike, and $\mu_R(x)$ is its triangular membership function.

Figure 3 illustrates the graphic presentation of the triangular membership function. The closer the value of preference/dislike degree to a Kano's standard

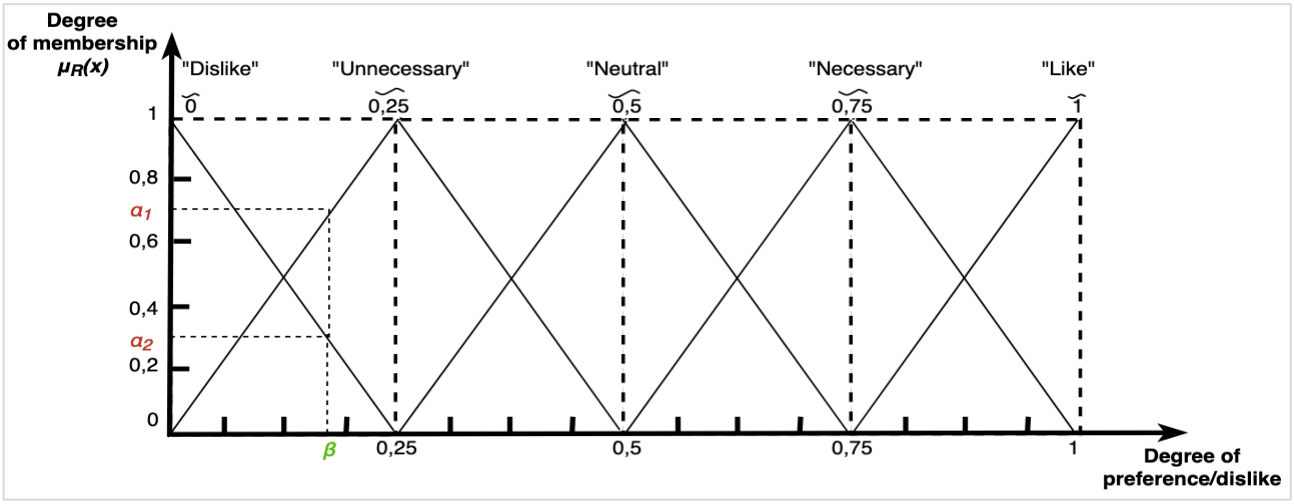


Figure 3 The triangular membership function of the degree of preference/dislike to the Kano standard answers.

answers, the higher the membership degree to it. For instance, while a *preference* value is located between 0 and 0.25, namely β, the membership degrees to "dislike" and "unnecessary" are α_1 and α_2 respectively.

In Table 2, we illustrate an example of a customer's membership degrees of preference and dislike for aspect 1 in topic 0. Using Table 2 only, it is difficult to determine the proper classification of the customer requirements. Therefore, the customer's membership degrees of preference and dislike can be transformed into two five-vector representations, namely $Pre = \{0.75, 0.21, 0.04, 0, 0\}$ and $Dis = \{0, 0, 0, 0.91, 0.09\}$ as defined in (Lee and Huang, 2009). Then, using a matrix multiplication $Pre^T \otimes Dis$, a 5×5 Kano's two-dimensional Fuzzy relation matrix '*MS*' is obtained as:

$$MS = Pre^T \otimes Dis = \begin{bmatrix} 0 & 0 & 0 & 0.68 & 0.06 \\ 0 & 0 & 0 & 0.19 & 0.01 \\ 0 & 0 & 0 & 0.03 & 0.003 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (6)$$

Relative to Table 1 stated in the literature, the customer requirements can also be written as a two-dimensional 5×5 matrix '*ME*' as:

$$ME = \begin{bmatrix} Q & A & A & A & O \\ R & I & I & I & M \\ R & I & I & I & M \\ R & I & I & I & M \\ R & R & R & R & Q \end{bmatrix} \quad (7)$$

After 'MS' being obtained, we sum the values of the 'MS' matrix entries with each other if they belong to the same cell in the evaluation matrix 'ME'. As a result, the classification of the customer requirements can be acquired as follows:

$$R = \left\{\frac{0.68}{A}, \frac{0.013}{M}, \frac{0.06}{O}, \frac{0.22}{I}, \frac{0}{R}, \frac{0}{Q}\right\} \quad (8)$$

As mentioned earlier, the Kano model's classification of requirements is qualitative and judged to be ineffective in the quantitative evaluation of customer satisfaction. Therefore, Berger et al. (1993) proposed customer satisfaction coefficients to provide quantitative values of satisfaction and dissatisfaction in case of fulfillment or non-fulfillment of a customer requirement, as given in equations 9 and 10:

$$CS_i^+ = \frac{A_i + O_i}{A_i + O_i + M_i + I_i} \quad (9)$$

$$CD_i^- = -\frac{O_i + M_i}{A_i + O_i + M_i + I_i} \quad (10)$$

Table 2 An example of a customer's membership degree to Kano's standard answers for aspect 1 in Topic 0. S = standard answers; M = membership degrees; Nec = necessary; Neu = neutral; Unnec = unnecessary; Dis = dislike.

S / M	Like	Nec	Neu	Unnec	Dis
Preference	75%	21%	4%		
Dislike				91%	9%

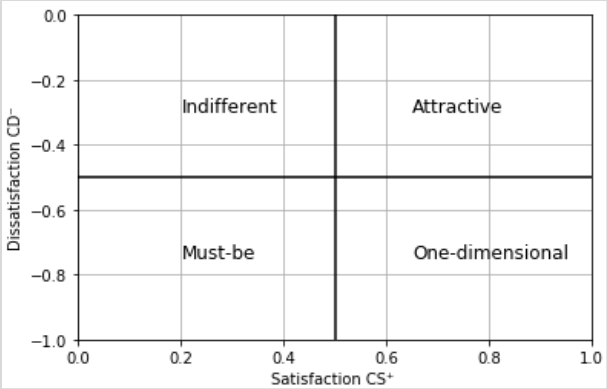


Figure 4 The Kano requirements classification according to customer satisfaction coefficients.

Where $CS_i^+$and CD_i^- are respectively the customer satisfaction and dissatisfaction coefficients of the i^{th} customer requirements, and A_i,O_i,M_i and I_i represent the probability distributions obtained according to the Kano's evaluation for the requirement i. Reverse and questionable requirements were ignored. Note that the minus sign in equation 10 emphasizes the negative impact on customer satisfaction, which will be decreased if these (one-dimensional and must-be) requirements are not included. On the other hand, the value of CS_i^+ is usually positive, indicating that customer satisfaction will be increased by providing these (attractive and one-dimensional) requirements.

A positive satisfaction coefficient ranges from 0 to 1, while a negative satisfaction coefficient runs from 0 to -1. A value of zero implies no impact on customer satisfaction whether the requirement is met or not. The closer CS_i^+ is to 1, the higher the influence of meeting the requirement is on the customer satisfaction, and the closer CD_i^- is to -1, the greater the influence of not meeting the requirement is on the customer dissatisfaction. In this way, all evaluated requirements can be represented graphically through a scatterplot, which is divided into four quadrants according to the satisfaction coefficient values. The X-axis is for CS^+ and the Y-axis is for CD^-. Each customer requirement could be assigned to different quadrants of the scatterplot based on the Kano requirements. As shown in Figure 4, the first quadrant stands for the one-dimensional requirements, the second quadrant stands for the attractive requirements, the third quadrant stands for the indifferent requirements and the fourth quadrant stands for the must-be requirements. Therefore, in designing new products/services, priority should be given to the higher CS^+ and the lower CD^- i.e. Attractive requirements, and when improving an existing product/service, more focus should be given to the high CS^+ value and the high CD^-value, i.e. one-dimensional requirements. This rule guides the decision-maker's team of a company when deciding on which customer requirement has more impact on the company's quality production process.

3.4 Decision Making Analysis driven by Fuzzy-Kano and SWOT

In this module, we propose a bi-layered matrix that maps the Fuzzy-Kano outputs into the SWOT matrix in order to interpret the requirements from the customer and the provider perspectives, as shown in Figure 5. The upper matrix lists the requirements from the customer's perspective. Its horizontal axis represents the fulfillment level of a requirement deducted from the customer satisfaction and dissatisfaction coefficients previously calculated, while the other axis refers to the Fuzzy-Kano requirement's classification. The upper matrix results are mapped into the SWOT matrix (lower matrix). SWOT is used as an analysis tool to provide insights about products by identifying their strengths and weaknesses (i.e. internal factors) along with potential opportunities and threats (i.e. external factors) (Phadermrod et al., 2019).

As can be seen from Figure 5, the upper matrix includes six zones ranging from (a) to (f). Zone (a) contains unfulfilled must-be requirements. The product's provider needs to fulfill these requirements in order to guarantee the minimum quality of the product. Zone (b) includes fulfilled must-be requirements which

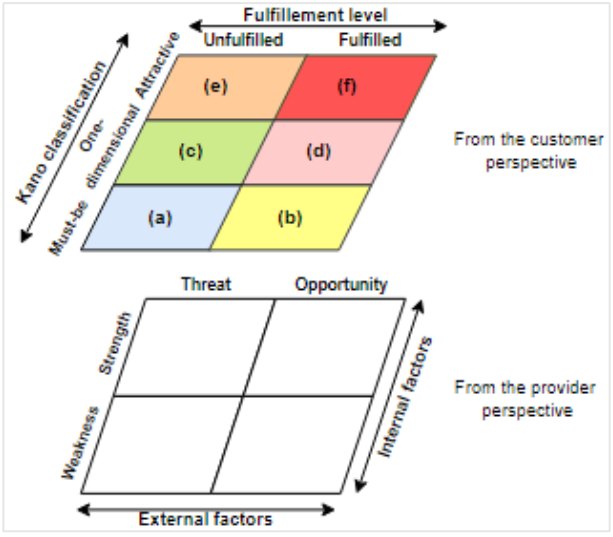


Figure 5 The KANO and SWOT bi-layered matrix.

means that the product already retains a minimum of quality. Zone (c) includes unfulfilled one-dimensional requirements. The product's provider should invest more in improving these requirements in order to avoid customer dissatisfaction and increase customer satisfaction. Zone (e) contains unfulfilled attractive requirements. Even though these requirements will not cause the customer dissatisfaction since they are not expected by the customers, they create a product with a novel attractive aspect that can achieve unexpectedly positive effects. Zones (d) and (f) hold fulfilled/one-dimensional and fulfilled/attractive requirements, respectively. The product's provider does not need to modify the product since those requirements are already at a high level of satisfaction. However, if they make more effective improvements, this can dramatically raise customer satisfaction. The improvements to be made in both zones are different. In (f), improvements are more innovative, while in (d) they are more realistic.

In the lower matrix, the aforementioned zones are mapped to the SWOT matrix. Zones (a) and (c) include unfulfilled/must-be and unfulfilled/one-dimensional requirements which can be regarded as a weakness of the product or even a potential threat for the provider. Therefore, zones (a) and (c) can be put in the W-T cell. Zone (e) holds unfulfilled attractive requirements that can be interpreted differently depending on the studied case. They can be considered as weaknesses that the product's provider can minimize by improving further the product quality and turn those weaknesses into an opportunity. In this case, zone (e) can be put in the W-O cell. On the other hand, those requirements can be considered strengths if the provider includes them in the product and they were not expected by the customers. However, if these requirements do not meet the customers' expectations, then they can become a potential threat. In this case, zone (e) can be put in the S-T cell. Zones (b), (d), and (f) respectively include the fulfilled/must-be, fulfilled/one-dimensional, and fulfilled/ attractive requirements that can be considered strengths since they can be easily fulfilled. In addition, adding new features to the product can be an opportunity to create a new market related to these features. Thus, these zones are put in the S-O cell.

Note that the indifferent requirements are not considered in the bi-layered matrix, simply because they are of little or no consequence to the customer. So, the provider can ignore them to save time, cost, and resources.

4. EXPERIMENTS AND RESULTS

In this section, we conduct a case study to evaluate the effectiveness and feasibility of the proposed framework using online mobile phone reviews collected from Amazon. In the following, we describe our dataset and show potential results.

4.1 Dataset

4.1.1 Preprocessing

In order to evaluate the effectiveness and feasibility of the proposed framework, the first phase consists of collecting and preprocessing the required dataset. In this paper, a dataset of unlocked mobile phone reviews has been selected. This dataset was acquired from Amazon using ("PromptCloud"). It includes 400,000 mobile phone reviews, containing product and customer information, ratings and plaintext reviews. In this study, we conducted the experiments on a subsample of the original dataset, which contains approximately 2000 reviews.

Table 3 Partial demonstration of experimental dataset.

Review	Price	Rating
I feel so LUCKY to have found this used (phone to us & not used hard at all), phone on line from someone who upgraded and sold this one. My Son liked his old one that finally fell apart after 2.5+ yea...	199.99	5.0
It's battery life is great. It's very responsive to touch. The only issue is that sometimes the screen goes black and you have to press the top button several times to get the screen to re-illuminate.	199.99	3.0

Table 3 illustrates some samples from the dataset. Each single review includes a considerable amount of unnecessary data, which must be cleaned to reduce noisy data and extract insightful information such as aspects and sentiments. The preprocessing operations applied in this work include tokenization, stop word removal, transform cases, stemming, and non-alphanumeric character removal. All the preprocessing operations were conducted using the Python NLTK toolkit (version 3.7). In addition, we grouped synonyms to reduce dimensionality by using a manually entered list including the most common synonyms e.g. the words "cellphone", "smartphone", "phones" are all transformed into "phone". Negation

```
Topic: 0 Word: 0.021*"bad" + 0.018*"worth" + 0.016*"turn" + 0.016*"purchase" + 0.013*"plastic" + 0.012*"problem" + 0.012*"dat
a" + 0.012*"may" + 0.011*"products" + 0.011*"battery" + 0.010*"could" + 0.010*"quality" + 0.010*"overall" + 0.010*"model" +
0.010*"problems" + 0.010*"mobile" + 0.010*"using" + 0.009*"price" + 0.009*"running" + 0.009*"loves"

Topic: 1 Word: 0.029*"super" + 0.022*"screen" + 0.020*"fast" + 0.018*"works_fine" + 0.018*"without" + 0.018*"love" + 0.018*"b
ig" + 0.017*"ive" + 0.016*"thin" + 0.015*"box" + 0.015*"us" + 0.014*"dont_NEG" + 0.014*"battery_NEG" + 0.014*"battery_life" +
0.014*"great" + 0.013*"performance" + 0.013*"cant" + 0.013*"away" + 0.012*"happy" + 0.012*"little"

Topic: 2 Word: 0.064*"nice" + 0.023*"unlocked" + 0.020*"well" + 0.019*"good" + 0.018*"times" + 0.017*"bought" + 0.016*"home"
+ 0.015*"grm" + 0.015*"calls" + 0.014*"likes" + 0.014*"husband" + 0.014*"work" + 0.014*"Verizon" + 0.013*"seems" + 0.013*"fee
l" + 0.012*"nd" + 0.012*"android" + 0.012*"like" + 0.011*"person_NEG" + 0.011*"money"

Topic: 3 Word: 0.027*"far" + 0.025*"got" + 0.024*"sim_card" + 0.023*"great" + 0.022*"around" + 0.022*"havent" + 0.020*"androi
d" + 0.018*"update" + 0.018*"amazing" + 0.017*"work_NEG" + 0.017*"straight" + 0.016*"doesnt" + 0.016*"wasnt" + 0.015*"connect
ed" + 0.015*"available" + 0.015*"ok" + 0.014*"constantly" + 0.014*"week" + 0.014*"ago" + 0.014*"never"
```

Figure 6 List of top 20 keywords for the first four topics.

handling is quite important in this study, it assists in improving sentiment analysis accuracy. Therefore, we used the simplest approach proposed in (Das et al., 2001), which is based on appending a negation tag *"_NEG"* to every word found between a negation and the first punctuation mark following it, so as to reverse the polarity of all these words while computing their scores. Misspelling is also taken into consideration since the reviews are usually hand-typed. Some predefined functions from the *"autocorrect package"* are used to deal with misspellings. The POS tagging is used to find adjectives that are considered sentiment words, as well as products' aspects where nouns (NN) and noun phrases (NNP) are considered potential aspect candidates.

Table 4 Setting values for running LDA.

Parameter settings	Values
Number of documents (M)	1593
Number of topics (K)	20
Number of iterations	50
$\alpha = 1/K$	1/20
$\beta = 1/K$	1/20

Table 5 List of aspects along with their sentiment polarity and scores for topic ID = 5.

Aspect(s)	Polarity	Sentiment score
Battery safety	-1	-0.72
Booting time	-1	-0.14
Price	1	0.53
Speakers quality	1	0.83
Battery life	-1	-0.57
Shipping	1	0.33
Screen size	-1	-0.92
Internet speed	-1	-0.10
weight	1	0.69
Camera resolution	1	0.86

Moreover, we applied certain filtering operations, such as: excluding reviews without an adjective POS tag, since sentiments are mainly identified from adjectives; pruning words that are not recognized by the opinion lexicon or Wordnet; and keeping reviews in which an aspect appeared at least once. In the end, the final list was made up of 1763 reviews, which was split into 1593 reviews intended for training and 170 reviews for testing. The testing reviews were chosen randomly, and a new column was added, including aspects and the relative sentiments' polarity.

4.1.2 Extracting Topics and Constructing Aspect-Sentiment Pairs

Before proceeding with the LDA application, we prepared the data for phrase modeling, which consisted of grouping common words that often get a special meaning when they are used together. That is, we built bi-gram phrases from the reviews. Then, using the "*GENSIM*" library, we built our LDA model over the parameters cited in Table 4. The number of topics K was set at 20 to avoid producing a general result with a lack of details. Moreover, a larger number of topics may take longer to converge. For the other parameters, *GENSIM* default values were used.

Through the LDA model, we obtained the first output, namely, the word-topic matrix. It included 20 meaningful topics each represented as a weighted list of words in descending order. Figure 6 indicates the first four topics with the top 20 most frequent words. Topics were inspected by a specific index. Instead, topic names can be defined manually by inferring topics from relevant words' meanings. For instance, looking at topic 1 keywords, we can summarize it to "phone screen and battery performance". The second output generated by LDA was the document-

topic matrix. An example of topic allocation to the five first documents (reviews) is illustrated in Figure 7.

By extracting numerous aspects that customers are reviewing and their corresponding sentiments along with the accumulated sentiment scores calculated using equation 2, we gain insights into what negatively or positively impacts product reviews, as well as what the customers like or dislike about the product. Table 5 shows a partial list of such aspects along with their polarity classes and sentiment scores grouping by topic ID 5.

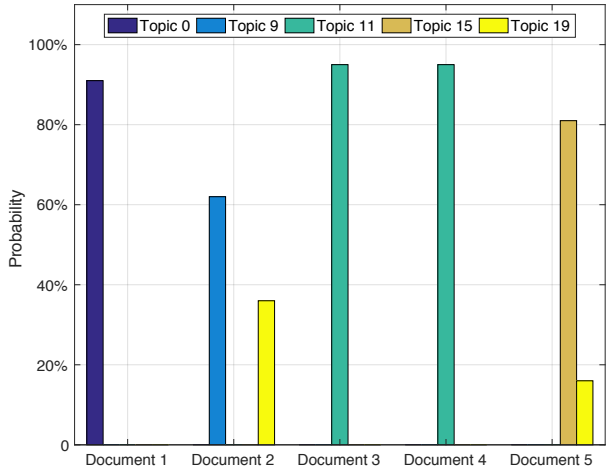


Figure 7 Topic distribution for the first 5 documents.

4.2 Evaluation and Results

4.2.1 Results of the Extracting Aspect-Sentiment Pairs

To evaluate how the extracting aspect-sentiment pairs approach performed, two set of experiments were conducted: (i) measure the effectiveness of the aspects extraction and (ii) measure the effectiveness of the sentiments assignment to the corrected aspects extracted. In this regard, four performance metrics were used: accuracy (Acc), precision (P), recall (R), and F1-score (F_1). Accuracy means how often our model is correct but when used alone, it cannot be trusted to select a well-performing model. Therefore, we used the three other metrics to give more detailed insights into the performance characteristics of our method. Precision refers to the percentage of the relevant data. A higher precision indicates more true positives and less false positives. On the other hand, recall expresses the proportion of all relevant results correctly classified by our model. High recall means less false negatives and high true positives. According to the confusion matrix notations (Ting, 2017), the accuracy, precision, and recall are computed respectively by the following equations:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$R = \frac{TP}{TP + FN} \tag{13}$$

Where TP is true positives, TN is true negatives, FP is false positives, FN is false negatives. The F1-score combines precision and recall and gives an overall view of the accuracy of the approach. The F1-score is given by:

$$F_1 = 2 * \frac{P \times R}{P + R} \tag{14}$$

In the experiment set (i), TPs refer to the correctly extracted aspects. TNs are the aspects that were discarded by the model and did not appear in the test data either. FPs are words that the model classified as aspects but are not actually aspects. FNs are the aspects that the model labeled as not being aspects when they were actually aspects. In the experiment set (ii), TPs refer to the aspects correctly classified with positive scores. FPs are the aspects incorrectly classified with positive scores. FNs are the aspects incorrectly classified with negative scores.

Table 6 Performance results. Acc = accuracy; Pre = precision.

Experiments set	Acc.	Pre	Recall	F1-score
(i) Aspects extraction	97.4%	92.4 %	84.5%	88.27%
(ii) Sentiments assignment	89,8%	90.7%	94.7%	92.6%

Table 6 depicts the accuracy, precision, recall, and F1-score of the proposed aspect-sentiment pairs approach in the experiments set (i) and (ii). As one can see, in (i), the model reports a high precision value (92.4%) meaning that most of the actual aspects are correctly classified with low FP values. The recall rate is 84.5%, suggesting that the most returned aspects are correctly labeled with low FN values. The F1-score is relatively high, meaning that the model represents insightful

results in terms of extracting the most discussed aspects of specific products. In (ii), the results are significantly different than the first experiment set. In particular, the F1-score is 92.6%, which indicates that assigning correct sentiments' polarity performs fairly well compared to the aspects' extraction, which reports 88.27%. These results suggest that the extraction of aspect-sentiment pairs performs efficiently in identifying accurate aspects and assigning appropriate sentiments to them. This will help in feeding the Fuzzy-Kano model with accurate inputs, consequently providing valuable business insights.

4.2.2 Results of the Fuzzy-Kano Model

The Fuzzy-Kano model classified the ten aspects previously extracted into must-be, one-dimensional, attractive, and indifferent requirements by calculating their degrees of preference and dislike. Table 7 highlights the findings of the assessed requirements' classification along with their impact on customer satisfaction.

According to the customer satisfaction coefficient (CS^+/CD^-) reported in Table 7, we can represent all the classified requirements via a scatterplot, as shown in Figure 8.

Table 7 Fuzzy-Kano classification and customer satisfaction coefficients results. R.No. = requirement number; A. Req. = assessed requirements; Kano Class = Kano Classification.

R. No.	A. Req.	Kano Class	CS^+	CD^-
R_0	Battery safety	Must-be	0.29	-0.83
R_1	Booting time	One-dimensional	0.78	-0.62
R_2	Price	Indifferent	0.06	-0.05
R_3	Speakers quality	One-dimensional	0.54	-0.58
R_4	Battery life	Must-be	0.46	-0.89
R_5	Shipping	Indifferent	0.42	-0.12
R_6	Screen size	Attractive	0.83	-0.36
R_7	Internet speed	One-dimensional	0.60	-0.70
R_8	Weight	Attractive	0.57	-0.32
R_9	Camera resolution	Attractive	0.71	-0.49

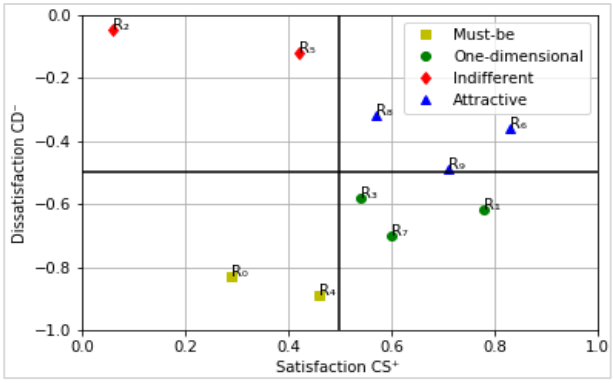


Figure 8 The representation of the Fuzzy-Kano classification results according to CS+ and CD-.

From Figure 8 and Table 7, the findings indicate that all the must-be requirements are battery-related, namely, R_0 and R_4 since they have a higher level of dissatisfaction among the customers compared to other requirements. Furthermore, R_1, R_3, and R_7 are all one-dimensional requirements, which implies that customers expect the companies to improve the performance of this product requirement. On the other hand, the attractive requirements such as R_6 and R_9 have a greater impact on satisfaction if fulfilled while R_8 has a relatively lower impact on customer satisfaction when compared to R_1. The indifferent attributes, R_2 and R_5 reflect a low impact on customer satisfaction and dissatisfaction, thus, they should be the last to be focused on over the three other requirements.

4.2.3 Fuzzy-Kano and SWOT Mapping and Analysis Results

In this section, the identified requirements are mapped to the bi-layered matrix. First, they are classified according to the Fuzzy-Kano model from the customer's perspective, then, classified according to the SWOT method from the provider's perspective. The results of the mapping are shown in Figure 9.

Considering the aforementioned results and the analysis reported in the fourth module of our proposed framework, R_0 and R_4 must be fulfilled to guarantee the minimum quality of the product and meet the customers' requirements. These requirements are headed to W-T, which motivate the provider to improve the battery performance, including safety and durability. In addition, internet speed (R_7) is considered W-O from the provider's perspective. Therefore, further enhancements of R_7 will not only lead to increased customer satisfaction but also decrease its dissatisfaction. Requirements in the zones (d)

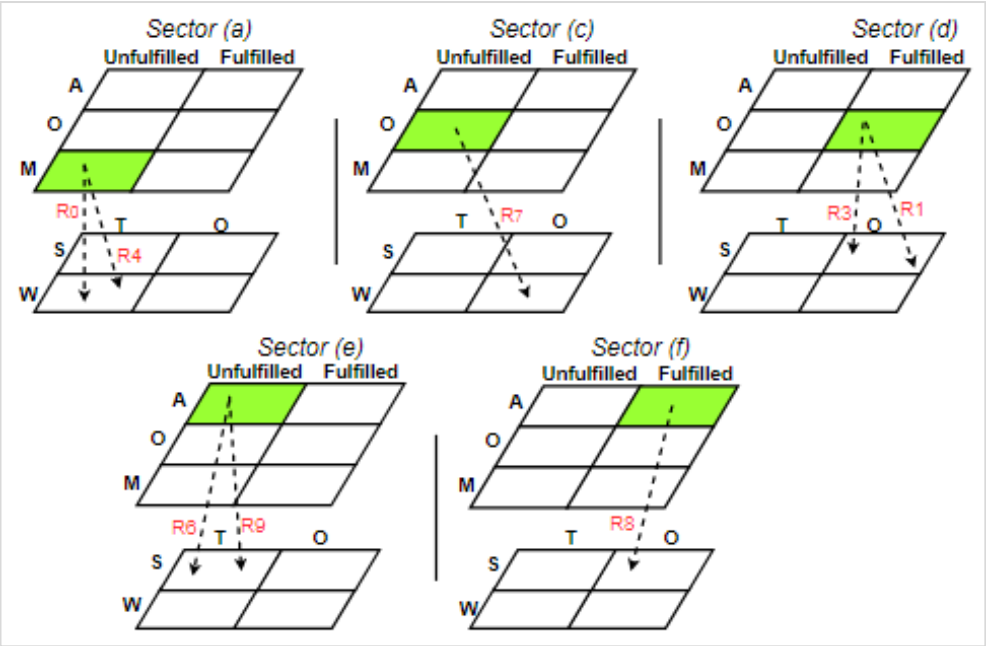


Figure 9 Requirements mapping results.

and (f) such as booting time (R_1), loudspeaker quality (R_3), and weight (R_8) are included in S-O, which means that those requirements are easy to fulfill, and when the provider makes more improvements on them, this will lead to a higher level of customer satisfaction than the current level. The requirements in zone (e) are related to S-T. Even though (R_9) and (R_6) are not expected by the customers, the provider should be able to assess the customers' preferences and overcome the current threat by adding a new value to the product, e.g. improve the camera resolution.

5. CONCLUSION

A good understanding of customer satisfaction is important for the survival of any company in today's competitive market. No business can deny the critical role of the customers' voices in increasing customer satisfaction. However, drawing insights from a huge amount of VOC data is challenging. Thus, companies resort to BI methods and tools to extract actionable information for improving their products and meeting their customers' needs.

This study proposes a decision-making framework for assisting companies in understanding their customers' satisfaction through extracting meaningful insights from online VOC data. The proposed framework consists of four main modules: data extraction and preprocessing, aspect-sentiment pairs extraction using LDA, requirement classification based on the Fuzzy-Kano model, and decision-making analysis driven by Fuzzy-Kano and SWOT.

A case study including online reviews of mobile phones is considered to evaluate the performance of the aspect-sentiment pair extraction module based on several metrics including the accuracy, precision, recall, and F-score. The results showed that the aspects were correctly extracted with a value of 97.4% in accuracy and 92.4 % in precision. Additionally, the sentiments were accurately assigned to the extracted aspects with a value of 89.8% and a precision value of 90.7%. These results constitute an accurate VOC input to feed the Fuzzy-Kano model. They allow us to classify the customer requirements that affect their satisfaction into four main categories: must-be, one-dimensional, attractive, and indifferent. Then, we can map them dynamically to the SWOT matrix in order to provide valuable and interpretable insights for companies.

This framework has some potential limitations that serve as a direction for future work. First, the study is conducted on online reviews which are assumed to be hand-typed and written by honest reviewers (i.e. not fake). However, if these reviews have been maliciously manipulated, they may impact the analysis process and result in biased decisions. An efficient spam review detection technique would be needed to identify whether the reviews are real or fake.

In addition, the aspect-sentiment pairs extraction module deals only with the explicit aspects but does not tackle the implicit ones. For example, in the following sentence *"The battery of this phone is pretty good"*, the aspect *"battery"* appears explicitly. However, in the

sentence *"The phone lasts all day",* the aspect *"battery"* is implicit because it is not stated directly, but only inferred from the meaning of the sentence.

Furthermore, the dynamics of the Fuzzy-Kano model are not included. It considers the evolution of the customer requirements over time. e.g., current attractive requirements can be transformed into must-be requirements in the coming years.

6. REFERENCES

Aguwa, C.C., Monplaisir, L., Turgut, O., 2012. Voice of the customer: Customer satisfaction ratio based analysis. Expert Systems with Applications 39, 10112–10119. https://doi.org/10.1016/j.eswa.2012.02.071

Alghamdi, R., Alfalqi, K., 2015. A survey of topic modeling in text mining. Int. J. Adv. Comput. Sci. Appl.(IJACSA) 6.

Berger, C.C., Blauth, R.E., Boger, D., 1993. kano's methods for understanding customer-defined quality.

Blei, D.M., 2012. Probabilistic Topic Models. Commun. ACM 55, 77–84. https://doi.org/10.1145/2133806.2133826

Carulli, M., Bordegoni, M., Cugini, U., 2013. An approach for capturing the Voice of the Customer based on Virtual Prototyping. J Intell Manuf 24, 887–903. https://doi.org/10.1007/s10845-012-0662-5

Culotta, A., Cutler, J., 2016. Mining Brand Perceptions from Twitter Social Networks. Marketing Science 35, 343–362. https://doi.org/10.1287/mksc.2015.0968

Darling, W.M., 2011. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 642–647.

Das, S.R., Chen, M.Y., Agarwal, T.V., Brooks, C., Chan, Y., Gibson, D., Leinweber, D., Martinez-jerez, A., Raghubir, P., Rajagopalan, S., Ranade, A., Rubinstein, M., Tufano, P., 2001. Yahoo! for amazon: Sentiment extraction from small talk on the web, in: 8th Asia Pacific Finance Association Annual Conference.

Decker, R., Trusov, M., 2010. Estimating aggregate consumer preferences from online product reviews. International Journal of Research in Marketing 27, 293–307. https://doi.org/10.1016/j.ijresmar.2010.09.001

Farhadloo, M., Patterson, R.A., Rolland, E., 2016. Modeling customer satisfaction from unstructured data using a Bayesian approach. Decision Support Systems 90, 1–11. https://doi.org/10.1016/j.dss.2016.06.010

Farhadloo, M., Rolland, E., 2013. Multi-Class Sentiment Analysis with Clustering and Score Representation, in: 2013 IEEE 13th International Conference on Data Mining Workshops. Presented at the 2013 IEEE 13th International Conference on Data Mining Workshops, pp. 904–912. https://doi.org/10.1109/ICDMW.2013.63

Gioti, H., Ponis, S.T., Panayiotou, N., 2018. Social business intelligence: Review and research directions. Journal of Intelligence Studies in Business 8.

Goodman, J., 2014. Customer experience 3.0: High-profit strategies in the age of techno service. Amacom.

Guo, Y., Barnes, S.J., Jia, Q., 2017. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. Tourism Management 59, 467–483. https://doi.org/10.1016/j.tourman.2016.09.009

Hofmann, T., 2017. Probabilistic Latent Semantic Indexing. SIGIR Forum 51, 211–218. https://doi.org/10.1145/3130348.3130370

Hu, M., Liu, B., 2004a. Mining Opinion Features in Customer Reviews, in: AAAI.

Hu, M., Liu, B., 2004b. Mining and Summarizing Customer Reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04. ACM, New York, NY, USA, pp. 168–177. https://doi.org/10.1145/1014052.1014073

Jia, S.S., 2018. Leisure Motivation and Satisfaction: A Text Mining of Yoga Centres, Yoga Consumers, and Their Interactions. Sustainability 10, 4458.

KANO, N., 1984. Attractive quality and must-be quality. Hinshitsu (Quality, the Journal of Japanese Society for Quality Control) 14, 39–48.

Lee, H., Han, J., Suh, Y., 2014. Gift or threat? An examination of voice of the customer: The case of MyStarbucksIdea. com. Electronic Commerce Research and Applications 13, 205–219.

Lee, Y.-C., Huang, S.-Y., 2009. A new fuzzy concept approach for Kano's model. Expert Systems with Applications 36, 4479–4484. https://doi.org/10.1016/j.eswa.2008.05.034

Lu, Y., Mei, Q., Zhai, C., 2011. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. Inf Retrieval 14, 178–203. https://doi.org/10.1007/s10791-010-9141-9

Miller, G.A., 1995. WordNet: a lexical database for English. Communications of the ACM 38, 39–41.

Nyblom, M., Behrami, J., Nikkilä, T., Solberg Søilen, K., 2012. An evaluation of Business Intelligence Software systems in SMEs-a case study. Journal of Intelligence Studies in Business 2, 51–57.

Park, Y., Lee, S., 2011. How to design and utilize online customer center to support new product concept generation. Expert Systems with Applications 38, 10638–10647. https://doi.org/10.1016/j.eswa.2011.02.125

Phadermrod, B., Crowder, R.M., Wills, G.B., 2019. Importance-Performance Analysis based SWOT analysis. International Journal of Information Management 44, 194–203. https://doi.org/10.1016/j.ijinfomgt.2016.03.009

PromptCloud: Fully Managed Web Scraping Service, n.d. URL https://www.promptcloud.com/ (accessed 9.24.19).

Qi, J., Zhang, Z., Jeon, S., Zhou, Y., 2016. Mining customer requirements from online reviews: A product improvement perspective. Information & Management, Big Data Commerce 53, 951–963. https://doi.org/10.1016/j.im.2016.06.002

Rese, A., Sänn, A., Homfeldt, F., 2015. Customer integration and voice–of–customer methods in the German automotive industry. International Journal of Automotive Technology and Management.

Reyes, G., 2016. Understanding non response rates: insights from 600,000 opinion surveys.

Sabanovic, A., Søilen, K.S., 2012. Customers' Expectations and Needs in the Business Intelligence Software Market. Journal of Intelligence Studies in Business 2.

Saura, J.R., Palos-Sanchez, P., Grilo, A., 2019. Detecting indicators for startup business success: Sentiment analysis using text data mining. Sustainability 11, 917.

Søilen, K.S., Tontini, G., Aagerup, U., 2017. The perception of useful information derived from Twitter: A survey of professionals. Journal of Intelligence Studies in Business, 7(3).

Szolnoki, G., Hoffmann, D., 2013. Online, face-to-face and telephone surveys—Comparing different sampling methods in wine consumer research. Wine Economics and Policy 2, 57–66. https://doi.org/10.1016/j.wep.2013.10.001

Ting, K.M., 2017. Confusion Matrix, in: Sammut, C., Webb, G.I. (Eds.), Encyclopedia of Machine Learning and Data Mining. Springer US, Boston, MA, pp. 260–260. https://doi.org/10.1007/978-1-4899-7687-1_50

Tirunillai, S., Tellis, G.J., 2014. Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. Journal of Marketing Research 51, 463–479. https://doi.org/10.1509/jmr.12.0106

Tontini, G., Solberg Søilen, K., Silveira, A., 2013. How interactions of service attributes affect customer satisfaction: A study of the Kano model's attributes. Total Quality Management & Business Excellence 24, 1253–1271.

Ullah, A.M.M.S., Tamaki, J., 2011. Analysis of Kano-model-based customer needs for product development. Systems Engineering 14, 154–172. https://doi.org/10.1002/sys.20168

Umoh, U.A., Isong, B.E., 2013. Fuzzy logic based decision making for customer loyalty analysis and relationship management. International Journal on Computer Science and Engineering 5, 919.

Xiao, S., Wei, C.-P., Dong, M., 2016. Crowd intelligence: Analyzing online product reviews for preference measurement. Information & Management 53, 169–182. https://doi.org/10.1016/j.im.2015.09.010

Xu, X., Li, Y., 2016. The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach. International Journal of Hospitality Management 55, 57–69. https://doi.org/10.1016/j.ijhm.2016.03.003

Journal of Intelligence Studies in Business
Vol. 9, No. 2 (2019) pp. 59-71
Open Access: Freely available at: https://ojs.hh.se/

A new corpus-based convolutional neural network for big data text analytics

Wedjdane Nahili[a]*, Kahled Rezeg[a] and Okba Kazar[a]

[a]*LINFI Laboratory, Computer Science, Biskra University, Algeria*

Corresponding author (*): w.nahili@univ-biskra.dz

Received 3 September 2019 Accepted 27 October 2019

ABSTRACT Companies market their services and products on social media platforms with today's easy access to the internet. As result, they receive feedback and reviews from their users directly on their social media sites. Reading every text is time-consuming and resource-demanding. With access to technology-based solutions, analyzing the sentiment of all these texts gives companies an overview of how positive or negative users are on specific subjects will minimize losses. In this paper, we propose a deep learning approach to perform sentiment analysis on reviews using a convolutional neural network model, because that they have proven remarkable results for text classification. We validate our convolutional neural network model using large-scale data sets: IMDB movie reviews and Reuters data sets with a final accuracy score of ~86% for both data sets.

KEYWORDS Convolutional neural networks, deep learning, natural language processing, NLP, user reviews, sentiment analysis, text classification

1. INTRODUCTION

The main purpose of sentiment analysis is analyzing and understanding expressed human emotion in text data. People are sharing daily thoughts and opinions about everything, and as a result, social media platforms have become the source of varied data, such as reviews of products, movies, and services. With the availability of this content a new type of information is harvested. Understanding 'what people think' and the real meaning of this user-generated data is crucial. Movie review sites such as IMDB, Rotten Tomatoes and Netflix represent an important source of information for researchers. The main reason behind this attention is the fact that valuable knowledge is often hidden behind this content and cannot be easily processed, which has gained increasing popularity among natural language processing (NLP) researchers. Deep learning algorithms are useful when it comes to solving natural language processing problems, and the reason resides in the combination of a large sample of data and a general learning algorithm (Collobert et al., 2011). Several methods can do this with traditional algorithms such as Naive Bayes and Support Vector Machine (SVM). Most of these methods consider the text word by word and classify a sentence as positive or negative by analyzing the words in the text. Sometimes information can be lost by extracting a keyword without another word (Shen et al., 2014). Recently, sentiment analysis research successfully used deep learning. Convolutional neural networks is one of the machine learning models that has archived remarkable results in image recognition and in natural language processing (Collobert et al., 2011).

In order to propose a text classification approach using deep learning, this work

introduces a new convolutional neural network architecture for text classification, solving different natural language processing tasks, specifically sentiment analysis. Our model's strengths are its training time and accuracy. In our sentiment analysis model, we utilize convolutional neural networks because they have impressive results in image analysis and classification fields. With their convolution operation they can extract an area of features from global information, and are able to consider the relationship among these features (Y. Kim, 2014). For computer vision, such as image analysis, convolutional neural networks are able to extract pixel data information. This means they can not only extract the pixels one by one, but also the feature information can be extracted piece by piece, where the piece contains multi-pixel data information. Thus, according to (Krizhevsky et al., 2012) when text is transferred into a matrix, it can also be considered to be the same as an image-pixels matrix. As a result, we can do the same operation to the text data to make the input features to the model that can be trained in another effective way (Yoon Kim, 2014).

In this paper, we propose a convolutional neural network (CNN) model to apply sentiment analysis on movie review data in order to predict sentiment orientation. Firstly, as an input to our network model, we use the word2vec proposed by Google to compute vector representations of words and reflect the distance between them. This step leads to initializing the parameters for our CNN model, therefore, efficiently improving the network performance in this particular problem. Secondly, we propose a CNN architecture with three convolution layers with padding, a flatten layer followed by two dense layers. To the best of our knowledge, using this layer architecture in a CNN model with an embedding layer (word2vec) to analyze movie reviews sentiment has not been addressed before in the literature. And finally, to improve the accuracy of our model, we use normalization and dropout layers.

The present work is organized as follows: Section 2 presents a brief literature background with some related concepts used in our approach. Section 3 outlines the related work on sentiment analysis and text classification, with an emphasis on deep learning methods. In Section 4, we present our approach and provide the description for the proposed architecture. In Section 5, the results and experimental setup are explained in detail along with the datasets used to train, test and validate our model and we present and elaborate on the performance using our model, and provide insight into the findings. Finally, we conclude our work and discuss future directions in Section 6.

2. BACKGROUND

2.1 Convolutional Neural Networks

Convolutional neural networks, also known as ConvNets, are a deep learning tool that has gained traction in computer vision applications (S. Srinivas et al., 2016). They were first introduced in Y. LeCun et al., (1989) to recognize handwritten ZIP code in 1989. They were later extended to recognize and classify various objects such as hand-written digits (MNIST), house numbers (P. Sermanet et al., 2012), Caltech-101 (L. Fei-Fei et al., 2007), traffic signs (P. Sermanet et al., 2011), and recently the work of A. Krizhevsky et al. (2012) produced a 1000-category ImageNet data set. The choice of using neural networks to create natural language processing (NLP) applications is attracting huge interest in the research community and they are systematically applied to all NLP tasks (Y. Kim, 2014).

The fundamental idea of CNNs is to consider feature extraction and classification as one joined task. The scope of using this methodology in text analytics has proven to be advantageous in various ways (D. Santos et al., 2014; A. Severyn et al., 2015; S. Srinivas et al., 2016). In deep learning techniques, there is supervised learning, unsupervised learning, hybrid learning and reinforced learning (A. Gibson and J. Patterson, 2017), but supervised learning and unsupervised learning are the most common techniques. The main difference is: in supervised learning, the data is labeled and known prior to training. This technique is suited for classification and regression problems. In unsupervised learning, the data is not labeled, which makes it good for clustering problem where algorithms can find different types of patterns within the unlabeled data (M. Mohri et al., 2012). With machine learning, there is deep structured learning, commonly known as deep learning. It can be used in different learning frameworks such as unsupervised, supervised and hybrid networks, in addition of different classification, regression and vision problems (L. Deng and D.Yu, 2014). A deep learning model can be described as a model of two nodes, where one is

an input, and the other an output. Data is sent between these two nodes through the input layer. The data is examined at different levels and features once it is sent onto the hidden layers.

Recently, CNNs have been adopted in natural language processing, sentiment analysis, text, topic and document classification for the following key reasons: CNN can extract an area of features from global information, it is able to consider the relationship among these features (Y. Kim et al., 2014), and text data features are extracted piece by piece and the relationship among these features, with the consideration of the whole sentence, thus, the sentiment can be understood correctly.

2.2 Sentiment Analysis

There are a number of different problems that deep learning is trying to solve. From classification problems where the algorithms assign categories to items, for instance, news categories, and to regression problems where the algorithm gives predictions on real values like a prediction on the stock market (M. Mohri et al., 2012). Another problem is sentiment analysis, also known as opinion mining. Sentiment analysis is an active research field in natural language processing, where people's emotions, opinions, and sentiments towards different entities like products, services, and organizations are studied and analyzed. Sentiment analysis is important for companies, organizations and individual persons (D. Tang, 2018). Companies want to know what people think about their products and services while on the other hand, individual people want to know what others think about a product they are considering purchasing. Daniel Angus stated: "This not only provides insight into what people think about your brand, but it can go a lot deeper. It can expose why people are thinking it."

In sentiment analysis, the goal is to determine whether a given piece of text is positive, negative or neutral. Various work has been done in the field of sentiment analysis in recent years where text is analyzed in several ways. In general, there are three levels of sentiment analysis: document-level, sentence-level and aspect-based level (A. Kharde, 2016).

Document-level: at this level, the analysis takes in consideration that the entire document has only one opinion.

Sentence-level: this level takes in consideration each sentence as containing one opinion and thus, the polarity of the entire document depends on the polarity of the sentences.

Aspect-based level: is also known as feature-based sentiment analysis. At this level, each sentence can contain more than one aspect in order to determine the polarity of the document (A. Kharde, 2016).

The main advantage of deep learning approaches in sentiment analysis remains in the fact that networks train themselves on the same data to learn the structures and context of the data. The data can vary and is often in the form of electronic data collected and made available for analysis. The crucial aspects of the data are the size and quality of the information. The better the quality of the data used in training, the better the results of predicting data in the future (J. Heaton, 2015).

2.3 Natural Language Processing

Natural language processing (NLP) is an industry term for algorithms designed to take a document consisting of symbols and deduce associated semantics (Russell. M, 2011). Research in NLP deals with the application of computational models to analyze text or speech data. Much work has been done in the field of NLP (Mikolov et al., 2013; Ouayang et al., 2015; Houshmand, 2017; Kalchbrenner et al., 2014) in order to allow semantic processing. Sentiment analysis is the research area where NLP algorithms are most often used, due to the amount of available data resulting from shared information on different social media platforms such as Facebook, Twitter, Amazon, Yelp, IMDB and Netflix. Until now, most sentiment analysis work has been done on short texts derived from social media sites. In this work, we analyze review texts because they provide sentiment about products or movies, therefore, when the result of this analysis is applied, it will help companies around the world to improve the decision-making process. Further, to automate sentiment analysis, different approaches have been applied to predict sentiments of words, expressions or documents (Mikolov et al., 2013; Ouayang et al., 2015; Houshmand, 2017; Kalchbrenner et al., 2014). These include NLP and deep learning methods. In our attempt to analyze the sentiment of movie review data and topic classification, we propose a deep learning approach that combines the advantages of available techniques such as CNNs along with NLP basic tasks. The following section reviews and discusses related work in the field of sentiment

analysis on reviews with emphasis on deep learning techniques.

3. RELATED WORK

Recently, much work has been done in the field of sentiment analysis in natural language and social network posts. To determine whether a piece of text expresses a positive or negative sentiment, two main approaches are commonly used: the lexicon-based approach and the machine learning-based approach. In recent years, deep learning models have achieved remarkable results in computer vision (Krizhevsky et al., 2012) and speech recognition (Graves et al., 2013). In the area of natural language processing, research on deep learning approaches (Bengio et al., 2003; Mikolov et al., 2013; Yih et al., 2011) has associated learning word vector representations. Although originally invented for computer vision and image analysis, CNNs have proven to be effective for NLP. These models have achieved impressive results in semantic parsing (Yih et al., 2014), search query retrieval (Shen et al., 2014), sentence modeling (Kalchbrenner et al., 2014), and various traditional NLP tasks (Collobert et al., 2011).

Ouayang et al. (2015) proposed a CNN and Word2Vec methodology for movie review sentiment analysis using a dataset from *rottentomatoes.com*. The data set contained 11,855 reviews with five different sentiment classifications (negative, somewhat negative, neutral, positive and somewhat positive). Their CNN model used three different convolution layers with different kernels and each layer was followed by a dropout layer and normalization layers. To evaluate their results, they compared their model against other algorithms/models including Naive Bayes, SVM, Recursive Neural Network (RNN) and Matrix-vector RNN (MV-RNN). The results show that performance is best when it comes to classifying every review into the five different classifications. Their model achieved a test accuracy of 45.4% on the test data set.

Houshmand (2017) compared different neural networks architectures against the Naive Bayes algorithm to see how well they performed on movie reviews from the Stanford Sentiment Tree bank dataset. The results of their study showed similar accuracy between the neural networks used (recurrent, recursive and convolutional neural networks) and Naive Bayes. One interesting thing about the result was the fact that their model's accuracy improved significantly by adding a word vector from Word2Vec to the network. Their model reached an accuracy of 46.4% on the test data while the CNN without a word vector had 40.5% accuracy (Table 1).

Table 1 Corpus-based related work.

	Corpus	**Accuracy**
Semantic parsing (Yih et al. 2014) CNN model		54%
Sentence modeling/sentiment analysis (Kalchbrenner et al. 2014) DCNN model	SST movie review TREC text retrieval	Binary class 86.8% Fine-grained 48.5%
Sentiment analysis (Ouayang et al. 2015) CNN+word2vec model	Rotten tomatoes movie review	Five classes 45.4%
Sentiment analysis (Houshmand, 2017) CNN model	STT movie reviews	40.5%
Sentiment analysis (Houshmand, 2017) CNN+word2vec model	STT movie reviews	46.4%

Despite the strong empirical performance in (Yih et al., 2014) and the good results in the work of (Mikolov et al., 2013; Ouayang et al., 2015; Houshmand, 2017; Kalchbrenner et al., 2014) we concluded that in (Yih et al., 2014) their system has no room for improvement because the corpus derived from the *WikiAnswers* data and *ReVerb KB* does not contain enough data to train a robust CNN model. Still, using word embeddings significantly improves the network's performance (Houshmand, 2017).

We propose a corpus-based CNN model to do sentiment analysis on a large-scale dataset (IMDB) in order to predict sentiment orientation. Firstly, similar to (Houshmand, 2017) as an input to our network model we use the word2vec as a lexical resource proposed by Google to compute vector representations of words and reflect the distance between them. This step leads to initialize the parameters at a good point of our CNN model. Secondly, the proposed sentiment analysis approach is done using a convolutional neural network architecture with three convolution layers with padding, a flatten layer followed by two dense layers with two dropout layers in between. To the best of our knowledge, using this architecture in a CNN model with an embedding layer to analyze movie reviews sentiment classification has not been addressed before in literature. Our results with

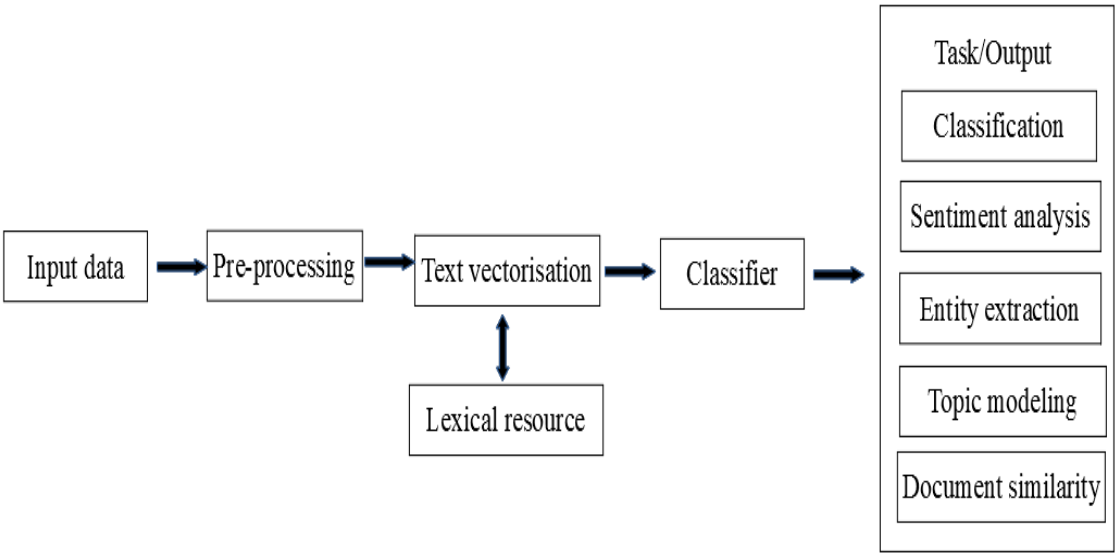


Figure 1 General architecture for text classification problems.

the proposed model have better results compared to related work.

4. PROPOSED APPROACH

With access to technology-based solutions and the rapid growth of social media platforms such as Twitter, Facebook, and online review sites such as IMDB, Amazon, and Yelp, users are sharing daily thoughts and opinions about different entities. These entities can be products, services, organizations, individuals, events, issues, or topics. This exponential growth of user-generated content draws growing attention from data scientists, as well as research and industry communities. The issue remains that reading every piece of this raw text data is time-consuming and resource demanding, therefore, analyzing this huge amount of text automatically gives companies an overview of how positive or negative users are to specific subjects will minimize losses. In order to automate this process work has been done in different fields like semantic parsing, sentence modeling and sentiment analysis (Mikolov et al., 2013; Yih et al., 2014; Ouayang et al., 2015; Houshmand, 2017; Kalchbrenner et al., 2014). Despite the results of previous work (Mikolov et al., 2013; Ouayang et al., 2015; Houshmand, 2017; Kalchbrenner et al., 2014), in addition to the strong empirical performance in Yih et al. (2014), their system has no room for improvement because the corpus does not contain enough data to train a robust CNN model. With the propose large-scale corpus-based model, we are able to obtain better results.

In this work, we use a CNN model to perform two tasks: binary-class sentiment analysis and multi-class text classification. In order to do so, first we analyze the sentiment of movie reviews using the publicly available IMDB dataset, then we classify news/ topics using the Reuters dataset. By using NLP, the computer can understand more than just the objective definitions of the words. This step includes using the word2vec model proposed by Google, which is a way of extracting features from the text for use in modeling, also using a classifier module to identify if a given piece of text is positive or negative in the case of sentiment analysis, and which topic or category the given piece of text fits into (Figure 1).

In this case, we are using a new CNN model as our classifier. Python libraries help the model learn with a faster curve, and the package “pandas” will help us read our CSV files containing both datasets. A Natural Language ToolKit (NLTK) is used to remove unnecessary data from the data sets. Figure 2 represents the process that takes place throughout the sentiment analysis process, which is divided into two sub-processes: the learning process where we train, test and validate our proposed CNN model and the classification process where new data is fed to the model. As illustrated in Figure 2, before any further analysis of the input text data, text pre-processing is needed, followed by text vectorization.

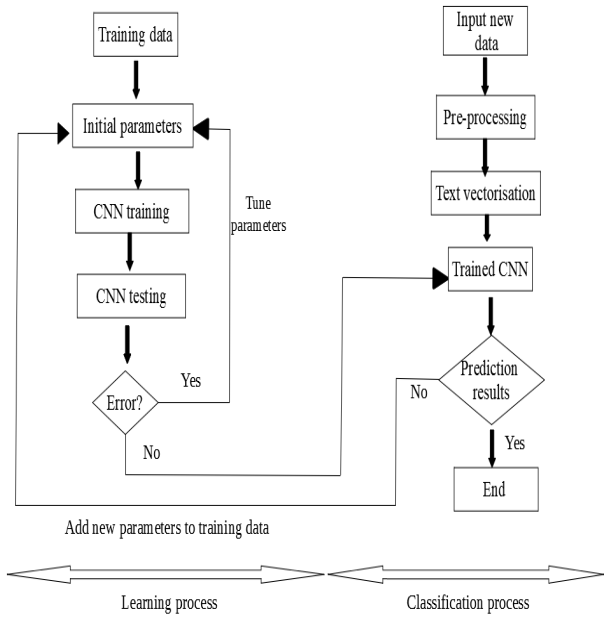


Figure 2 Global architecture for the proposed system.

4.1 Data pre-processing

It is necessary to normalize the text for any natural language processing tasks. Since it is often represented in a cryptic and informal way, systematic pre-processing of reviews is required to enhance the accuracy of our sentiment classifier. In this work, we perform a corpus-based analysis on text from users' movie reviews. Since natural language is frequently used in reviews, this type of text data contains a lot of noise as shown in Example 1, therefore, cleaning unnecessary information from raw comments (reviews) is needed. The movie review binary-class dataset used is IMDB, which contains 50,000 movie reviews labeled by sentiment (positive/negative). Similar to any NLP task, before any further processing, cleaning-up the data is crucial which involves the following steps:

1. Remove numeric and empty texts
2. Remove punctuation from texts
3. Convert words to lower case
4. Remove stop words

As demonstrated in Example 1, the datasets used contain non-relevant data (noise). Therefore, basic cleanup needs to be performed. Arbitrary characters and other useless information such as punctuation, stopwords, special characters and links/URLs were removed, since we found no significance in our classification approach. Then, text normalization was applied using regular expressions. When these NLP tasks are completed, the processed reviews are stored in a comma-separated value (CSV) file for further processing.

Stemming and lemmatization are text normalization (or sometimes called word normalization) techniques. This step is very important in order to get better accuracy for the proposed CNN model, and it consists of preparing the text, words, and documents for further processing. In order to stem and lemmatize words, sentences and documents, we used the public Python nltk package, the Natural Language Toolkit package, provided by Python for NLP tasks, as shown in Example 2.

Example 1:

[1] "I was blessed to have seen this movie last night. It made me laugh, it made me cry and it made me love life. This movie is a great movie that depicts a love of a father for his son. Will Smith did an incredible job and deserves every accolade available to him. His son also did a fantastic job. There is a great lesson that is learned in this movie and it truly shares the struggles of everyday life. This movie was heart felt and touching. It was truly an experience worth having. Thank you for making this movie and I look forward to seeing it again."

[1] "blessed night made laugh made cry made love life great depicts love father son incredible job deserves accolade son fantastic job great lesson learned shares struggles everyday life heart felt touching experience worth making forward"

Example 2:

"Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured,[1][2] similar to data mining."

4.2 Text vectorisation

In order to convert string features into numerical features, one can use one of the following methods.

One hot encoding maps each word to a unique ID, it has typical vocabulary sizes. They will vary between 10,000 and 250,000. This method is a natural representation to start with, though a poor one due to several drawbacks such as the size of input vector

scales with size of vocabulary. There is the "out-of-vocabulary" problem (H. L. Trieu et al., 2016) where there is no relationship between words (each word is an independent unit vector). Also it is vulnerable to overfitting: sparse vectors which result in computations going to zero (T. Ojeda et al., 2018).

Bag of words is an approach where we set all words in the corpus (T. Ojeda et al., 2018). Its main advantage is that it is quick and simple. But it is too simple and orderless, without syntactic or semantic similarity.

N-gram model is a model with a set of all n-grams in the corpus. It tries to incorporate the order of words (T. Ojeda et al., 2018), unfortunately it still has a very large vocabulary set and no notion of syntactic/semantic similarity.

Term frequency-inverse document frequency is a model that captures the importance of a word (term) to a document in a corpus. The importance of a word increases proportionally according to the number of times a word appears in the document; but is contrarily equivalent to the frequency of the word in the corpus (T. Ojeda et al., 2018). The key advantage of this method is that it is easy to compute and has some basic metric to extract the most descriptive terms in a document. Thus it can easily compute the similarity between two documents using it, but it does not capture the position in the text, semantics and co-occurrences in different documents because it is based on the bag-of-words model.

Thus term frequency-inverse document frequency is only useful as a lexical resource, but it cannot capture semantics like topic models and word embedding. In our work we use word2vec published by Google in 2013, which is a neural network implementation that learns distributed representations for words (Mikolov et al., 2013). Prior to word2vec, other deep or recurrent neural network architectures had been proposed (Ouayang et al., 2015; Kalchbrenner et al., 2014) for learning word representations. The major problem with previous attempts was the long time required to train the models, while word2vec learns quickly compared to these models. In order to create meaningful representations, word2Vec does not need labels. Since most data in the real world is unlabeled, this feature is very useful. If the network is trained on a large dataset, it produces word vectors with interesting characteristics. As a result, words with similar meanings appear in clusters, and clusters are spaced such that some word relationships, such as analogies, can be reproduced using vector math.

4.3 Convolutional Neural Network classifier

We propose a word-based CNN architecture for both binary-class and multi-class text classification. First, there is a sentiment analysis on the IMDB movie reviews dataset, which contains 50,000 movie reviews labeled by sentiment (positive/negative), and second a text (topic) categorization for the Reuters corpus, which contains 10,788 news documents totaling 1.3 million words, where the documents have been classified into 90 topics and grouped into two sets. As shown in Figure 3, we train a CNN with an embedding layer and different convolution layers with padding. The purpose of using padding in every convolution layer is to conserve the size of the input data as it is; thus, no information is lost (Shen et al., 2014). These convolution layers are followed by a flatten layer and two dense layers with two dropout layers.

4.3.1 Sentence matrix

Instead of image pixels, the input to most NLP tasks is sentences or documents represented as a matrix. Each row of the matrix corresponds to one token, typically a word, but it could be a character (Krizhevsky et al., 2012). That is, each row is a vector that represents a word. Typically, these vectors are word embeddings

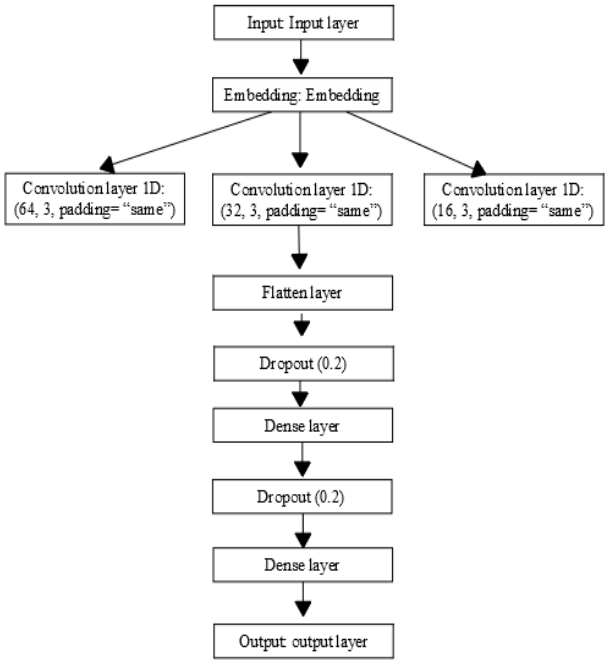


Figure 3 The layer architecture of the proposed CNN model.

like word2vec or Glove. For example in our work, a 10 word sentence using a 300-dimensional embedding, has a 10×300 matrix as input. That's our input sentence matrix (image) to the network (Y. Kim et al., 2014).

4.3.2 Embedding Layer

As input to our proposed model, the first layer is an embedding layer which is defined as the first hidden layer and its role is to transforms words into real-valued feature vectors known as embeddings. These vectors are able to capture morphological, syntactic and semantic information about the words. It must specify the following arguments: top-words, embedding-vector-length, and max-review-length. In this work, we truncate the reviews to a maximum length of 1600 words and we only consider the top 10,000 most frequently occurring words in the movie reviews dataset, and we used an embedding vector length of 300 dimensions. This is an important step in the proposed network architecture because it initializes the parameters of our CNN model.

The output of the embedding layer is a 2D vector (none, max-review-length, embedding-vector-length) with one embedding for each word in the input sequence of words. Some modification is applied to the basic convolutional operation (layer) where padding is used to conserve the original size of the input sentence matrix, therefore, there is no loss of information (Shen et al., 2014). To connect the dense layer (fully connected layer) to the 2D output matrix we must add a flatten layer in order to convert the output of the convolution layers into a single 1D vector to be used by the dense layer for final classification (Figure 4).

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 1600, 300)	3000000
conv1d_1 (Conv1D)	(None, 1600, 64)	57664
conv1d_2 (Conv1D)	(None, 1600, 32)	6176
conv1d_3 (Conv1D)	(None, 1600, 16)	1552
flatten_1 (Flatten)	(None, 25600)	0
dropout_1 (Dropout)	(None, 25600)	0
dense_1 (Dense)	(None, 180)	4608180
dropout_2 (Dropout)	(None, 180)	0
dense_2 (Dense)	(None, 1)	181

Total params: 7,673,753
Trainable params: 7,673,753
Non-trainable params: 0

Figure 4 Total number of trainable parameters in our CNN model.

4.3.3 Fully activated Layer (Dense)

In deep learning models, activation functions are used at the fully activated layer (dense) and they can be divided into two types: linear activation functions and non-linear activation functions (ML, 2018). In our work, the first experiment is binary-class sentiment analysis using the IMDB dataset where we used the sigmoid activation function. We used a sigmoid function because it exists between 0 to 1. Therefore, it is adequate for our model since we have to predict the probability as an output. In the second experiment we train, test and validate our CNN model on a multi-class Reuters dataset. We used the soft-max activation function since it is a more generalized logistic activation function, which is used for multi-class classification.

4.3.4 Dropout Layer

With approximately 7 million trainable parameters, the proposed CNN model is very powerful. However, overfitting is a serious problem in large networks, making them slow to use and thus difficult to deal with overfitting by combining many different predictions. Dropout is a technique that prevents this problem and it refers to dropping out units (hidden and visible) in a neural network (Lai. S-H et al., 2017). By dropping a unit out, we mean temporarily removing it from the network, along with all its incoming and outgoing connections. In our model we use two dropout layers with (0.2), and the choice of which units to drop is random.

5. RESULTS AND DISCUSSION

We propose a CNN model to apply text classification. We define a CNN model and we train it on publicly available data sets: th IMDB movies reviews dataset and the Reuters dataset. Our model is word-based CNN with an embedding layer. At the embedding layer level, we tokenize text review sentences to a sentence matrix with rows where each row contains word vector representations of each token. In our work, we truncate the reviews to a maximum length of 1600 words and we only consider the top 10,000 most frequently occurring words in the movie reviews dataset. We experiment with the network model in two settings. The first experiment involves predicting sentiment classification of movie reviews and the second one is news/topic

classification. The network performs well in both the binary and the multi-class experiments.

5.1 Datasets

As shown in Table 3, to evaluate the performance of our proposed model, we used two large scale datasets, the binary class IMDB dataset for sentiment classification (A. Maas et al., 2011) and the multi-class Reuters data set for news/topic classification (Table 2).

Table 2 IMDB and Reuters datasets.

IMDB	Reuters
#of sentences 50k	# of documents 10788
#of positive reviews 25k	# of topics 90
#of negative reviews 25k	# of word 1.3 million

We benchmark our CNN model on two different corpora from two different domains: movie reviews and news/topic classification. The movie review binary-class dataset used is IMDB, which contains 50,000 movie reviews labeled by sentiment (positive/negative). Reviews have been pre-processed, and each review is encoded as a sequence of word indexes (integers). This allows for quick filtering operations such as: "only consider the top 10,000 most common words, but eliminate the top 20 most common words" (A. Maas et al., 2011). In our experiments, we focus on sentiment prediction of complete sentences (reviews). The second corpus we use is the Reuters news wire topic classification. This dataset is a multi-class benchmark (e.g. there are multiple classes), multi-label (e.g. each document can belong to many classes) dataset (M. Thoma, 2018). Both datasets are used to validate our model, where the first dataset is the IMDB movies reviews. The data was split evenly with 25,000 reviews intended for training and 25,000 for testing. Moreover, each set has 12,500 positive and 12,500 negative reviews. We pre-processed the reviews, and each review is encoded as a sequence of word indexes (integers). And the second dataset is the Reuters dataset for document classification; it has 10,788 news documents and 90 classes/topics.

We conduct an empirical exploration on the use of the proposed word-based CNN architecture for sentiment classification on IMDB movie reviews and the Reuters corpus for text categorization, which contains 10,788 news documents totaling 1.3 million words where the documents have been classified into 90 topics and grouped into two sets. In the present work, we train a CNN with an embedding layer, convolution layers, a flatten layer and two dense layers with two dropouts. Although CNNs extract high-level features in image analysis, our model actually performs well in 2D problems and trains 50% to 60% faster as shown in Figures 5 and 6. The proposed model has ~7M trainable parameters and is trained in a Python environment which takes around 15 to 20 minutes on an Intel (R) Core (TM) i5-5200U CPU with 2.20GHz of RAM.

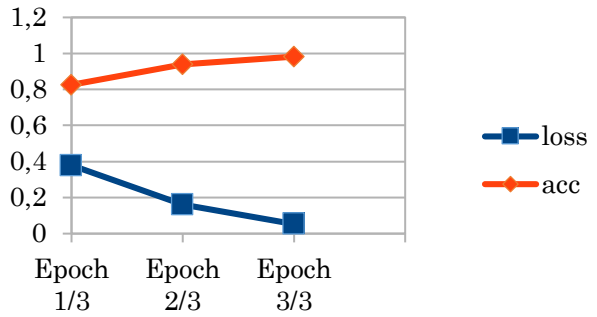


Figure 5 Loss function and accuracy values of the proposed model on the IMDB dataset.

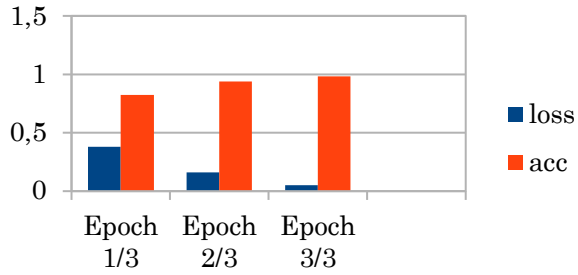


Figure 6 Loss function and accuracy values of the proposed model on the Reuters dataset.

In the sentiment classification of movie reviews using the IMDB dataset, in order to horizontally extract features, we used binary cross entropy loss because it is a binary classification problem. To avoid overfitting the training data dropout (0.2) was necessary. For reinforcing the generalization power, we disabled the network with holes during training. This way the network is forced to build new paths and extract new patterns. Despite the satisfactory performance of our model, and in addition we were able to validate the proposed model on both IMDB and Reuters datasets. After 15 to 20 minutes of training, we obtain ~86% accuracy (Table 3).

Table 3 Accuracy of the models on the IMDB dataset for binary-class and Reuters dataset for multi-class.

	Fine-grained	Binary
CNN model (Yih et al. 2014)		54%
DCNN model (Kalchbrenner et al. 2014)	48.5%	86.8%
CNN+word2vec model (Ouayang et al. 2015)	45.4%	
CNN model (Houshmand, 2017)		40.5%
CNN+word2vec model (Houshmand, 2017)		46.4%
CNN model	**85.95%**	**85.80%**
CNN+ LSTM model		**95%**

We tried to improve the accuracy of the model by conducting other experiments using a modified CNN and Long Short-Term Memory (LSTM) architecture. The embedding layer is still the first hidden layer of our CNN-LSTM model, we added the LSTM layer followed by GlobalMaxpool 1D layer, and 2 Dense layers with Dropout. The main difference between the CNN model and the CNN-LSTM model is at this level where we have the first dense layer with the 'ReLu' activation function instead of 'sigmoid' in the first CNN model. Similar to the experiments with our CNN model, in order to avoid overfitting, a dropout layer (0.5) was necessary. This layer is followed by the second dense layer where a 'sigmoid' activation function is used. The same NLP tasks are applied to the reviews which involve the following steps:

1. Remove numeric and empty texts
2. Remove punctuation from texts
3. Convert words to lower case
4. Remove stop words
5. Stemming

Only the IMDB dataset was used to train, test and validate the proposed CNN-LSTM model. The labeled dataset consists of 50,000 IMDB movie reviews, selected for sentiment analysis. The sentiment of reviews is binary, meaning the IMDB rating below 5 results in a sentiment score of 0, and ratings equal to or greater than 7 have a sentiment score of 1 and no individual movie has more than 30 reviews.

5.1.1 Raw Reviews

- *'With all this stuff going down at the moment...'*
- *'The Classic War of the Worlds by Timothy Hi...'*
- *'The film starts with a manager (Nicholas Bell)...'*
- *'it must be assumed that those who praised this...'*
- *'Superbly trashy and wondrously unpretentious 8...'*

5.1.2 Processed reviews

- *'stuff go moment mj ive start listen music watch...'*
- *'classic war world timothy hines entertain film...'*
- *'film start manager nicholas bell give welcome...'*
- *'must assume praise film great film opera ev...'*
- *'superbly trashy wondrously unpretentious 80 ex...'*

The 25,000 review labeled as the training set do not include any of the same movies as the 25,000 review test set. In addition, there are another 50,000 IMDB reviews provided without any rating labels.

The labeled training set is tab-delimited and has a header row followed by 25,000 rows containing an id, sentiment, and text for each review. The test set is a tab-delimited file that has a header row followed by 25,000 rows containing an ID and text for each review. The task of our CNN-LSTM model is to predict the sentiment for each. An extra training set with no labels is provided that is a tab-delimited file with a header row followed by 50,000 rows containing an ID and text for each review.

One interesting thing about the results of the CNN-LTSM model is that the accuracy improved significantly compared to the first CNN model. The CNN-LSTM model reached an F1 score of 0.95 on the test data while the

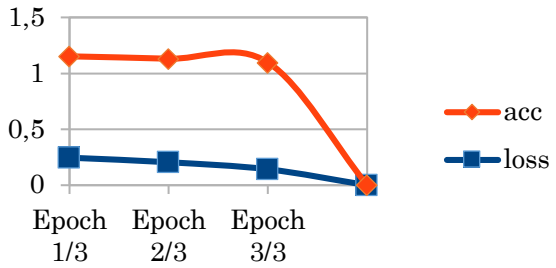


Figure 7 Loss function and accuracy values of the proposed CNN-LSTM model.

CNN without the LSTM layer got ~ 86% (Figure 7). We conclude that both models perform well and show satisfactory results against state-of-the-art methods, which is quite respectable given: (1) the large size of the data sets and (2) the number of parameters in the network.

6. CONCLUSION

With an aim of classifying the sentiment of movie reviews into two classes (positive or negative) and applying text classification on news text in order to perform topic classification, our method has been implemented with an acceptable performance. As a next step of making use of a data driven model, CNN has been taken into consideration. In this work we present a new CNN architecture that jointly uses word2vec as an input layer to the CNN model and an LSTM layer. The proposed model has yielded better results compared to previous methods with an accuracy of ~86 % for the first experiment and 95% for the CNN-LSTM (Mikolov et al., 2013; Ouayang et al., 2015; Houshmand, 2017; Kalchbrenner et al., 2014). The main contributions of the paper are: (1) the short training time despite the large size of the data sets and the number of parameters in the network; (2) the demonstration that adding an LSTM layer to the network can be effective and significantly improving the model's accuracy. In future research it will be interesting to apply the proposed model architecture to other NLP applications such as spam filtering and web searches, as well as exploring Bayesian optimization frameworks and also, conducting other experiments using recursive neural network with the long short-term memory architectures for sentiment categorization of text review.

7. REFERENCES

Bengio, Y. R. Ducharme, P. Vincent, and C. Jauvin, (2003). A Neural Probabilistic Language Model. Journal of Machine Learning Research, (3), 1137-1155.

Bing Liu, (2011). Opinion Mining and Sentiment Analysis, WEB DATA MINING. Data Centric Systems and Applications, Part 2, 459-526.

Bing Liu, (2012). Sentiment analysis and opinion mining. San Rafael, CA: Morgan and Claypool Publishers.

Britz, D. (2015). Understanding Convolutional neural networks for NLP, in WildML. Retrieved October 17th, 2018, from http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuglu, and P. Kuksa. (2011). Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, (12), 2493–2537

Deng, L. and D. Yu, (2014). Deep learning: Methods and applications. Grand Rapids, MI, United States: Now publishers.

Fei-Fei, L., R. Fergus, and P. Perona. (2007). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 objects categories. Journal of Computer Vision and Image Understanding, 106(1), 59-70.

Gibson, A. and J. Patterson, (2017). Deep Learning. Chapter 1: A review on machine learning. O'Reilly Media, Inc.

Graves, A. (2013). Generating sequences with Recurrent Neural Networks. Retrieved August 13th, 2018, from https://arxiv.org/abs/1308.0850

Heaton, J. (2015). Artificial intelligence for humans, volume 3: Deep learning and neural networks. United States: Createspace Independent Publishing Platform.

Houshmand, Shirani-Mehr, (2017). Applications of Deep Learning to Sentiment Analysis of Movie Reviews. Retrieved December 6th, 2018, from https://cs224d.stanford.edu/reports/Shirani-MehrH.pdf

Kalchbrenner, N., E. Grefenstette, and P. Blunsom. (2014). A Convolutional Neural Network for Modelling Sentences. In Proceedings of ACL 2014.

Kharde, A. and S. Sonawane, (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications, Volume 139, No.11, 0975-8887

Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), (pp. 1746–1751)

Krizhevsky, A., I. Sutskever, and G. Hinton, (2012). Imagenet classification with deep convolutional neural networks. In Advances in

neural information processing systems, 1097-1105

Lai, S-H., V. Lepetit, K. Nishino, and Y. Sato, (2017). Computer Vision – ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II, volume 10112, doi 10.1007/978-3-319-54184-6, 183-204

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. Journal of Neural Computation, 1(4), 541-551

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner. (1998). Gradient-based learning applied to document recognition. In proceeding of the IEEE, 86(11), (pp. 2278-2324).

Machine Learning Cheatsheet, (2018). Activation Functions. Retrieved December 6th, 2018, from https://ml-cheatsheet.readthedocs.io/en/latest/activation_functions.html

Maas, A. et al., (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, volume 1, (pp. 142- 150)

Micolov, T., K. Chen, G. Corrado, and J. Dean, (2013). Efficient Estimation of Word Representations in Vector Space. Journal of Computing Research Repository.

Mohri, M., A. Rostamizadeh, and A. Talwalkar, (2012). Foundations of machine learning. Cambridge, MA: MIT Press.

Ojeda, T., R. Bilbro and B. Bengfort, (2018). Applied Text Analysis with Python. Chapter 4. Text Vectorization and Transformation Pipelines. O'Reilly Media, Inc.

Ouyang, X., P. Zhou, C. H. Li, and L. Liu. (2015). Sentiment analysis using Convolutional neural network. In IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing.

Russell, M. (2011). Mining the social web, O'Reilly Media.

Santos, D., and C. Gatti, (2014). Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, (pp. 69–78)

Semanet, P., S. Chintala, and Y. LeCun. (2012). Convolutional neural networks applied to house numbers digit classification. In Proceeding of the 21st International Conference on Pattern Recognition (ICPR), (pp. 3288-3291).

Semanet, P., and Y. LeCun. (2011). Traffic sign recognition with multi-scale convolutional networks. In Proceeding of International Joint Conference on Neural Networks (IJCNN), (pp. 2809-2813).

Severyn, A., and A. Moschitti, (2015). Twitter sentiment analysis with deep convolutional neural networks. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, (pp. 959–962)

Shanmugamani, R., and R. Arumugam, (2018). Hands-On Natural Language Processing with Python. Packt Publishing.

Shen, Y., X. He, J. Gao, L. Deng, and G. Mesnil. (2014). Learning Semantic Representations Using Convolutional Neural Networks for Web Search. In Proceedings of WWW 2014.

Srinivas, S., R. Sarvadevabhatla, K. Mopuri, N. Prabhu, (2016). A taxonomy of deep convolutional neural nets for computer vision. Frontiers in Robotics and AI 2, 36

Tang, D., and M. Zhang, (2018). Deep Learning in Sentiment Analysis. In: Deng L., Liu Y. (eds) Deep Learning in Natural Language Processing. Springer, Singapore, 219-253

Thoma, M. (2017). The reuters dataset, Retrieved October 23rd, 2018, from https://martin-thoma.com/nlp-reuters/

Trieu, H.L., L. M. Nguyen and P. T. Nguyen, (2016). Dealing with Out-Of-Vocabulary Problem in Sentence Alignment Using Word Similarity. Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation (PACLIC 30). 259-266

Yadav, V. (2017). How neural networks learn nonlinear functions and classify linearly non-separable data?, Medium, Retrieved October 19th, 2018, from https://medium.com/@vivek.yadav/how-neural-networks-learn-nonlinear-functions-and-classify-linearly-non-separable-data-22328e7e5be1

Yih, W., K. Toutanova, J. Platt, and C. Meek. (2011). Learning Discriminative Projections for Text Similarity Measures. In Proceeding of the Fifteenth Conference on Computational Natural Language Learning CoNLL'11. (pp. 247-256).

Yih, W., X. He, and C. Meek. (2014). Semantic Parsing for Single-Relation Question answering. In ACL Proceeding.

Zhang, Y. and C. Wallace, (2016). A Sensitivity Analysis of Convolutional Neural Networks for Sentence Classification. Cornell University Library, Computer Science, Computation and Language

Journal of Intelligence Studies in Business
Vol. 9, No. 2 (2019) pp. 72-81
Open Access: Freely available at: https://ojs.hh.se/

Using open data and Google search data for competitive intelligence analysis

Jan Černý[a], Martin Potančok[a]* and Zdeněk Molnár[a]

[a]*Faculty of Informatics and Statistics, Department of Information Technologies, University of Economics, Prague, Czech Republic*

Corresponding author (*): martin.potancok@vse.cz

Received 23 September 2019 Accepted 28 October 2019

ABSTRACT Open data are information entities that are of significant importance for many institutions, businesses and even citizens as the part of the digital transformation within many fields in our society. The aim of this paper is to provide a competitive environment analysis method using open source intelligence within the pharmaceutical sector and to design the optimal data structure for this purpose. Firstly, we have described the state-of-the-art of open human medicine data within the European Union with a focus on antidepressants and we have chosen the Czech Republic as the primary research territory for demonstrating competitive intelligence analysis. Secondly, we have identified the competitive intelligence and open source intelligence relationship with a new possible contextual analysis method using open human medicine data and Google Search data. Finally, this paper shows the potential of open deep web data within competitive intelligence activities, together with surface web data entities as a low-cost approach with high intelligence value focused on the pharmaceutical market.

KEYWORDS Competitive intelligence, data structure, digital transformation, open data, open source intelligence, OSINT

1. INTRODUCTION

Open data plays a significant role in our present society and is one of the most important digital transformation trends. Moreover, it has become a solid part of the activities of business units that are charged with business analyses, insights and strategy plans (Janssen et al. 2012). The reason can be found in a very broad spectrum of industries and areas where open data has started to be a rational form of result output. As the number and scope of such open datasets has grown enormously to include in the areas of transportation, public services, natural science, education, demography, and last but not least the health sector, it has also become a significant part of many national information policies, shifting from governmental down to local levels. In the USA, a growing trackable significance was evident during the Obama administration after the official Data.gov site was launched (Kostkova et al. 2016). Data is also an essential part of the EU's Digital Single Market strategy, as "The EU needs to ensure that data flows across borders and sectors and disciplines. This data should be accessible and reusable by most stakeholders in an optimal way." (European Commission 2018). Moreover, massive digitalization and increasing information system/information and communication technology (IS/ICT) usage have brought big data challenges and demands for non-traditional analytical methods to uncover global and regional trends (Gandomi and Haider 2015).

This means, therefore, that open data appears to be a strong tool for a spectrum of competitive intelligence (CI) and open source intelligence (OSINT) methodologies at all levels of different industries and organization types. CI can be defined at its basic level as the process of planning, collecting and disseminating data, information and knowledge for the purpose of better decision-making, eliminating risks and uncovering of business opportunities, primarily in an external company environment (Grèzes 2015). The first phase identifies particular business information needs through key intelligence topics (KIT), and then key intelligence questions (KIQ) which analysts use for the collection process as they define information requests (Herring 1999). OSINT consists of a very similar cycle as CI, but it also consists of an open data and information source mining process through the collection phase. In addition, OSINT end-users do not come primarily from business, but from government, military, intelligence services and from the security service sector.

In the present paper we have focused on open human medicine data from two perspectives: government and business. In both cases we wanted to design open data intelligence analysis methods for the public sector, e.g. policy makers, ministers, commissioners and other key persons. Our governmental direction is directed towards setting up the optimal open data structure, and our business direction is directed towards complex business environment analysis. To demonstrate our intent, we have chosen open human medicine data focused on antidepressants. To increase specificity, we have added a Google Search data perspective to gain a territorial dimension for our analysis.

Our two main research questions are: could open data provide significant CI insights within the pharmaceutical industry? and could the surface web search data deliver a territorial perspective to anonymous open human medicine data?

2. LITERATURE REVIEW

There are several studies of how open data could help in the health sector. For example, Bernard et al. (2018) seek possible open source solutions that could be used for the detection, reporting and control of disease outbreaks, and analyze the previous use of similar tools in Ebola and SARS epidemics. One part of this work also considers the ethical level (Oubrich 2011) of these intelligence activities within the context of data ownership. This question is also discussed by Kostkova et al., (2016). Brownstein, et al. (2008) find the power of public information sources in the signal intelligence scope for outbreak-oriented detection activities to be at the local level of information sources such as discussion sites, disease reporting networks, and news outlets (with regards to a very detailed verification process). Google Search data played an important role in the past within the Google Flu Trends project. As Cook, et al. (2011) show in their evaluation, this tool was highly accurate in the prediction of influenza activity in the United States based on user search queries. Akhgar, et al. (2016) demonstrate the complex usage of OSINT methods, however the critical issue is also focused on an early warning system for health hazards. Open innovations (Hughes 2017) and open data (European Commission 2019) initiatives are also more visible in the health sector over recent years. For example, Cantor, et al. (2018) developed a dataset for community-level social determinants of health and strengthened the decision-making process for care planning. Farber (2017) discusses whether data repositories can help find effective treatments for complex diseases. His suggestions for informatics communities consist of methods concerning the provision of an effective data infrastructure with the inexpensive method of data accessibility from different datasets, monitoring the growth of biomedical datasets and finding ways to link data in different repositories.

Perer and Gotz (2013) and Hu, et al. (2016) have illustrated how data-informed and data-driven decisions can be supported by data visualization in the health sector. To achieve appropriate data visualization several principles should be followed (e.g. simplify, compare, explore) (Few 2012).

2.1 Survey and preparations

The quality of open datasets is an aspect of many discussions at all levels of the policy-making process. Policy makers put high pressure on the availability and frequency of the "update" aspect, but we would like to raise concern over the poor open data structure concept with regards to quality. Within this context, we highlight our recent study directed toward open human medicine data in the European Union (Cerny et al. 2018). During the first three months, we made a large survey

of national medicine control offices and their open data policies. The results showed significant differences.

Further, we have designed a method for the new CI approach and demonstrated the context for open human medicine data and Google Search data. There were a number of reasons for this step. To begin with, according to our secondary survey, five billion Google queries are conducted per day, and even a small sample of this amount could lead to significant insights. The second reason is that national control offices provide datasets strictly anonymously, with no territorial information. And, through the Google Trends application, we were able to mine the information-seeking behavior of surface web users and get the following data entities:

- The searcher interest rate
- The territorial origin of the searchers (region, city)
- Trend keywords connected to our desired terms

3. KEY INTELLIGENCE QUESTIONS

After we defined the research questions, we continued and narrowed our information needs through the following key intelligence questions (KIQ):

- KIQ1: Is antidepressant use increasing in the Czech Republic?
- KIQ2: What is the most prescribed antidepressant on the market?
- KIQ3: Who is the key player in the specific market?
- KIQ4: What is the market share of antidepressants on the specific market?
- KIQ5: How can Google Search data help to determine territorial information seeking behavior regarding antidepressant-oriented queries?

4. MATERIAL AND METHODS

4.1 Medicine data structure

We have analyzed data accessibility from national control agencies that are in charge of regulatory and distribution policies in the European Union, Switzerland, Norway and Turkey. When we contacted each agency, we collected information about time response, level of content relevancy feedback and the human factor based on their ability to help regarding the requested data collection. We were aware that the results from this primary research are strictly qualitative and could be misleading, so we broadened the timeframe of the research to three months. E-mail communication has been chosen as the first method of contact, however, in specific cases phone communication was also needed.

Secondly, we went through all the possible information sources, e.g. official websites, repositories and FTP servers, and monitored whether the open human health datasets are available and how they are handled with respect to their format. If the datasets did not exist online, we concentrated on a search system interface that could be used to generate datasets with the required data fields. If there was no evidence of the existence of open data, we contacted the person in charge of communication to gather information about the state of the open data policy. As our aim is to gain market insights regarding antidepressants, we have focused on the specific data entities that could lead to quality business analysis. Table 1 suggests the fields we have monitored and that, in our opinion, could uncover specific market trends. Here we explain why our suggested data fields should be considered to be key information elements for advanced CI business analysis.

Table 1 Designed data field for a complex business analysis.

ATC	Anatomical Therapeutic Chemical classification system
CODE_MED	Specific code of the medicine
NAME	Name of the medicine
ADDITIONAL_ INFORMATION	Additional information to the name
PRODUCER	Registration holder
COUNTRY_ORIGIN_ PRODUCER	Country of a registration holder
NUMBER_OF_ PACKAGES_YEAR	Number of packages / year
PRICE_NOSUR CHARGE_EXCL_VAT	Price per package excl. a surcharge and VAT
TOTAL_SUM_NOSUR CHARGE_EXCL_VAT	Total sum / all packages / excl. a surcharge and VAT
PRICE_SURCHARGE_ INCL_VAT	Price per package incl. a surcharge and VAT
TOTAL_SUM_SURCH ARGE_INCL_VAT	Total sum / all packages / incl. a surcharge and VAT
NUMBER_DDD	Defined daily doses / package
TOTAL_DDD	Defined daily doses / total
DDD_1000INH_DAY	Defined daily doses / 1000 inhabitants

Firstly, the ATC code (WHO 2018) is the internationally respected classification in the pharmaceutical field. We can demonstrate its role in our case study. As shown below, we have chosen the N06A group, but if the specific active component is needed for analysis, we could narrow it down and be more specific.

- **N** NERVOUS SYSTEM
- **N06** PSYCHOANALEPTICS
- **N06A** ANTIDEPRESSANTS
- **N06AA** Non-selective monoamine reuptake inhibitors
- **N06AB** Selective serotonin reuptake inhibitors
- **N06AF** Monoamine oxidase inhibitors, non-selective
- **N06AG** Monoamine oxidase A inhibitors
- **N06AX** Other antidepressants

The lowest level of the classification is further divided into specific medicines and this could be a crucial factor for resolving the situation when the datasets do not include commercial medicine names. For example, class N06AA (non-selective monoamine reuptake inhibitors) covers subclass N06AA01 (desipramine) along with information about the daily defined dose (DDD). In this scenario we would use MeSH Browser (U.S. National Library of Medicine 2019) to uncover commercial names, e.g. Pertofran, Norpramin among others. The specific code of the medicine supports ATC codes as the existence confirmation identifier of the specific medicine. The name of the medicine, its additional information and the producer, together with the country of origin, are the basic identifiers of any possible commercial entity analysis. The significance of the market activity of a given producer, or possibly of a specific medicine, uncovers the total number of prescribed packages with their total cost with no surcharge and excluding value added tax. Additional price fields are used for the price comparison of individual medicines.

4.2 Google data structure

Further, our intention was directed towards the process that could verify our open human medicine data CI market analysis results. If we were able to get detailed market data about pharmaceutical companies, we would also need to add territorial information, which is crucial because of the strict anonymity of open human medicine data. Through the Google Trends application data, we were able to mine the information-seeking behavior of surface web users and obtain the following data entities:

- The searcher interest rate with retrospectivity to the year 2004
- The territorial origin of the searchers (region, city)
- Trend keywords connected to our desired terms
- We have structured the Google Search data sets as follows:
- Country
- Search term (the keyword antidepressant in a given national language and in English)
- Week (in a specific year)
- Number of searches in given country
- Region
- Number of searches in given region

5. RESULTS

5.1 Open human medicine data analysis results

The data collected reflect the present level of open human health data quality and accessibility. We went through all three levels of the collection process and found significant differences. The biggest issues we faced could be identified as the different data structure in each of the countries together with language barriers leading to difficulties as to when data should be used in a whole-region analysis. Some datasets were complex (e.g. the Czech Republic and Slovenia), while others provided only simple insights into specific medicines, e.g. Wales or Slovakia, and others, e.g. Bulgaria or Greece, had no data. Although some countries had neither open repositories nor data files accessible, a few of them did provide a specific search interface that could be used for searching, filtering and exporting open medicine data. This approach is advantageous because the exported files already include the requested class of the medicine. Excluding France, we could define the classes in the search forms with the specific ATC code. Poland, Croatia and Lithuania especially have powerful search interfaces.

The third level of the data collection phase found significant differences between the information services of the agencies. Table 2 summarizes the response time, with comments. Where references are mentioned, the agency provided links to repositories, or to search interfaces.

Table 2 Survey of agency information service time response. R = days until response.

Entity	R	Institution Remarks
Austria	1	AGES (1 day), BASQ did not respond
Bulgaria	5	BDA (no data availability)
Croatia	-	HALMED (no response)
Czech Republic	1	SÚKL (reference to the Czech datasets)
Estonia	5	REAM (did not provide datasets with requested fields)
Finland	-	FIMEA (no response)
France	-	AMELI (no response)
Hungary	20	OGYÉI (references)
Germany	7	Several institutions contacted. Only paid datasets
Italy	-	AIFA (no response)
Latvia	5	ZVA (reference to the search interface)
Malta	3	Medicine Authority Malta (limited data availability)
Netherlands	4	CBG-MEB (do not provide requested data publicly)
Norway	7	NORPD (references)
Poland	5	URPL (no requested data availability)
Portugal	17	Infarmed (provided data only for study purposes)
Romania	5	ANM (references)
Turkey	6	TITCK (references)
Slovakia	2	ŠÚKL (only paid datasets)
Slovenia	-	JAZMP (no response)
Spain	3	AEMPS (no cooperation)
Sweden	15	LMF – Läkemedelsverket (references)
Switzerland	2	Interpharma (only limited data)
United Kingdom	18	MHRA (contacted several times, references)

During the collection process, we dealt mainly with data structure and data quality obstacles and did not get relevant support for our open data CI analysis research. The file formats and structure field values were different in every country. Moreover, the data quality implied high time costs in preparation for data analysis, especially when we dealt with the company and medicine name differences in each of the analyzed countries.

For the purpose of this paper we have chosen the open data CI analysis possibilities in the dataset from the Czech Republic. Firstly, the Czech dataset structure and quality was the most complex of the monitored countries and it is a great example of what can be achieved by open data. Secondly, we were able to make valuable market insights, even though the complexity of the data from the Czech Republic provided a powerful example of what can be achieved regarding competitive business intelligence. However, then we added the comparison possibility between the states with less structured content to demonstrate the minimum analysis context. The requested class of medicine was antidepressants, according to research question two. We have used Tableau (2019) to create an interactive visualization which can be shared and analyzed (Datig and Whiting 2018).

By focusing on the Czech Republic, we can gain very detailed insights. To begin with, we wanted to analyze the antidepressant consumption trend among Czech citizens (KIQ1). We used an open dataset covering the time period from 1991 to 2018. Figure 1 demonstrates the increase in antidepressant consumption during this period.

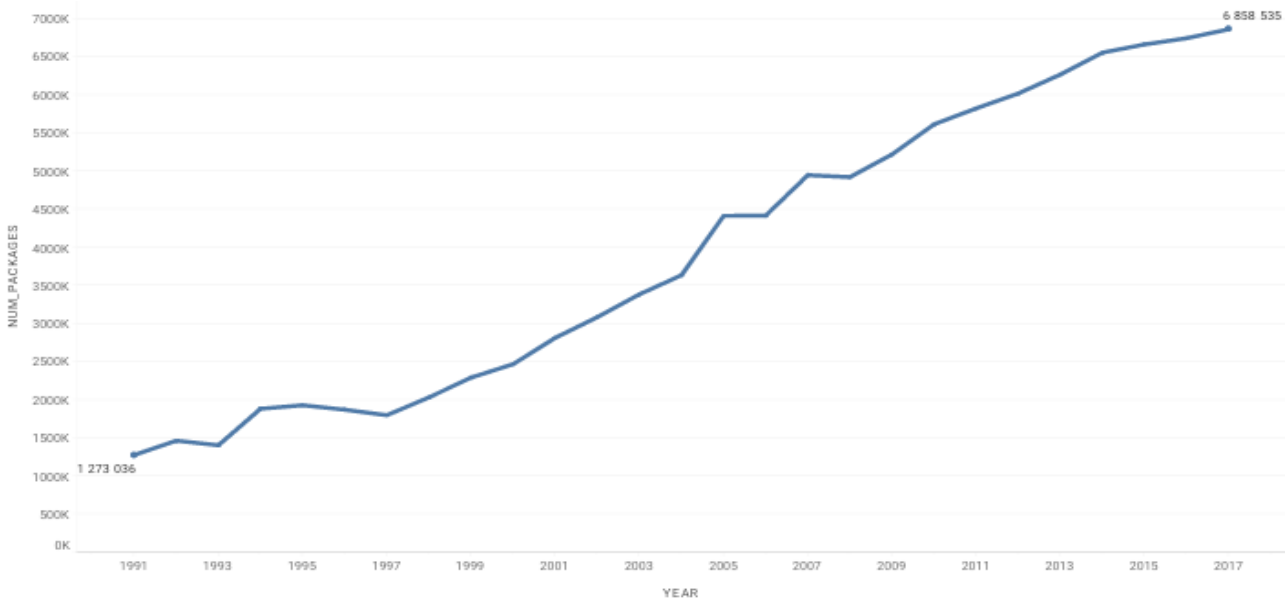


Figure 1 Czech antidepressant consumption 1991-2018.

Table 3 Czech antidepressant medicine leader market insights through open data 2009-2018 with prescribing information.

Producer	Medicine Name	Total packages	Percent packages	Treatment (from prescribing information)
H. Lundbeck	Cipralex	5 891 747	23.29	Depression and anxiety disorders (panic disorder with or without agoraphobia, social anxiety disorder, generalized anxiety disorder and obsessive-compulsive disorder)
Zentiva	Citalec	4 976 308	22.19	Depression and anxiety disorders
Pfizer	Zoloft	3 881 554	15.34	Depression with or without anxiety, panic disorder and obsessive-compulsive disorder and the treatment of post-traumatic stress disorder
Krka	Asentra	3 843 125	15.19	Depression and prevention of depression (adult), social anxiety disorder (in adults), post-traumatic stress disorder (in adults), panic disorder (in adults), obsessive compulsive disorder (in adults and children and adolescents aged 6-17)
Angelini	Trittico	3 763 352	14.88	Anti-anxiety, tension, restlessness, sleep disturbance, and sexual function

In the context of this trend, our further point of interest was to uncover the most prescribed antidepressants (KIQ2, KIQ3) and their market share in the country. We have narrowed the time period, as demonstrated in Table 3, to get the most accurate market data. This step was necessary due to significant market changes, e.g. Prothiaden was de facto the most prescribed antidepressant medicine until the year 2005, and then its popularity fell rapidly.

As we can uncover the main medicine representatives in the Czech Republic, we can connect these with the types of the mental disorder as shown in Table 3. This is used to predict mental health trends in a particular area.

However, thanks to open data, we can monitor the whole market share of antidepressants (KIQ4) and compare whether the producer of the main medicine representatives is similar to the whole antidepressant market share (Table 4).

Table 4 Czech antidepressant market share 2009-2018.

Producer	Total antidepressant packages	Percent of all packages sold
Zentiva	12 193 957	19,46
Krka	9 838 079	15,70
H. Lundbeck	7 680 389	12,25
Angelini	4 006 332	6,39
Pfizer	3 974 158	6,34

5.2 Google search data results

To analyze the context between open human medicine data and information seeking behavior we have used Google Trends (Nuti et al. 2014) including Google Search data (the context description is given above). The aim of the analysis was to confirm a correlation between Google Search data and market information about specific antidepressant consumption. Google Trends (available on trends.google.com) providing Google Search data in an available form and as confirmed by (Nuti et al. 2014) and (Nuti et al. 2014) is used by the health sector.

The identified set of related keywords (general antidepressant terms and specific names of the medicine) was gradually inserted into Google Trends, all data was downloaded into a CSV file and aggregated. It is important to emphasize that data was downloaded at the regional level for the period of analysis. Consolidated CSV files were used as a basis for the following analysis (Figure 2).

The conversion per capita was used for the analysis to ensure comparable results between countries with different population sizes. The overall analysis done in number of searches per capita shows an increase in searches since 2011 and confirms the increase in consumption based on the analysis above. As the analysis shows, the relationship between the number of searches and the number of searches per capita is not affected only by size of the country, but also by other factors. Norway, Estonia, Switzerland, Netherlands and Austria are among the countries with the largest number of searches per capita. Both number of searches

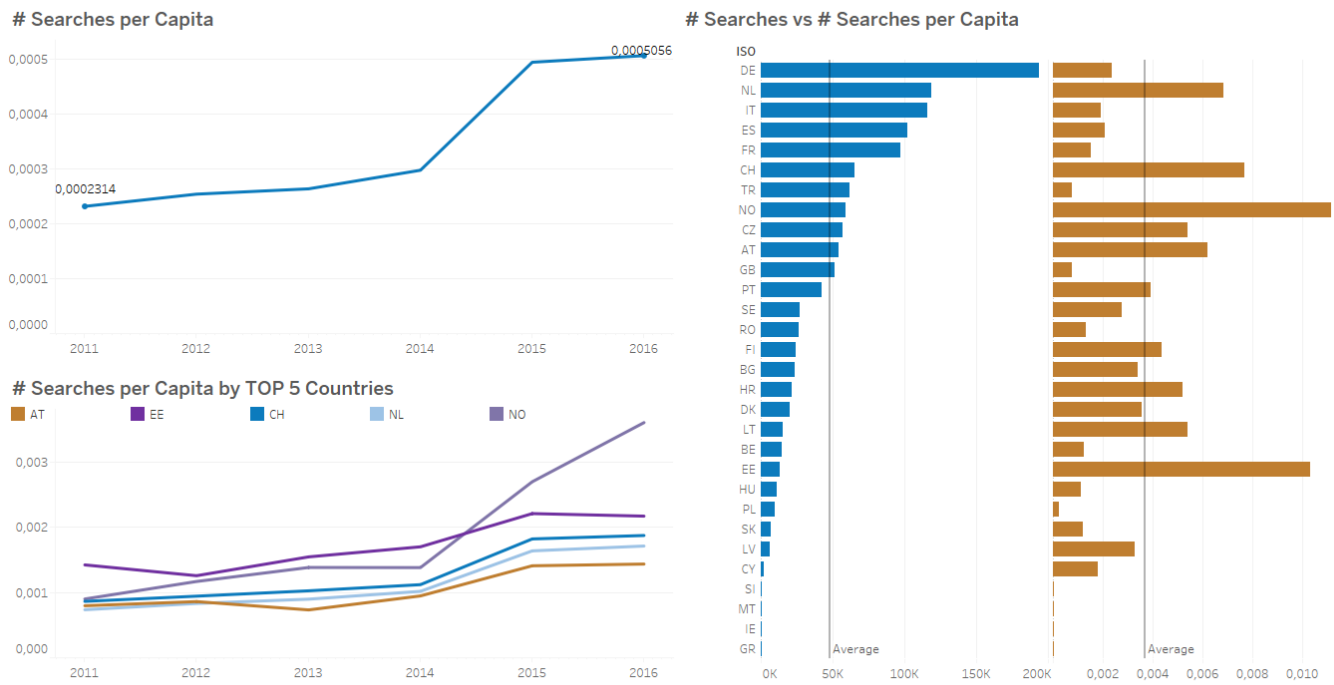


Figure 2 Search per capita analysis.

and number of searches per capita are above average in the Czech Republic.

It is important to analyze the correlation between Czech market trends and Czech searching trends. We have chosen the two most prescribed medicines in the Czech Republic, Cipralex and Citalec, and compared their market performance with Google search performance (Figure 3 and Figure 4).

We demonstrate with these analyses that there is a significant similarity between market data and Google data. To sum up, if the state control office does not provide open human medicine data with territorial dimension, we can use Google data (KIQ5) to narrow our market analysis (Table 5 and Table 6).

6. CONCLUSIONS

Open human health data can be considered to be crucial information entities for competitive environment analysis and for showing particular health trends across a large geographic area. There are two conditions that

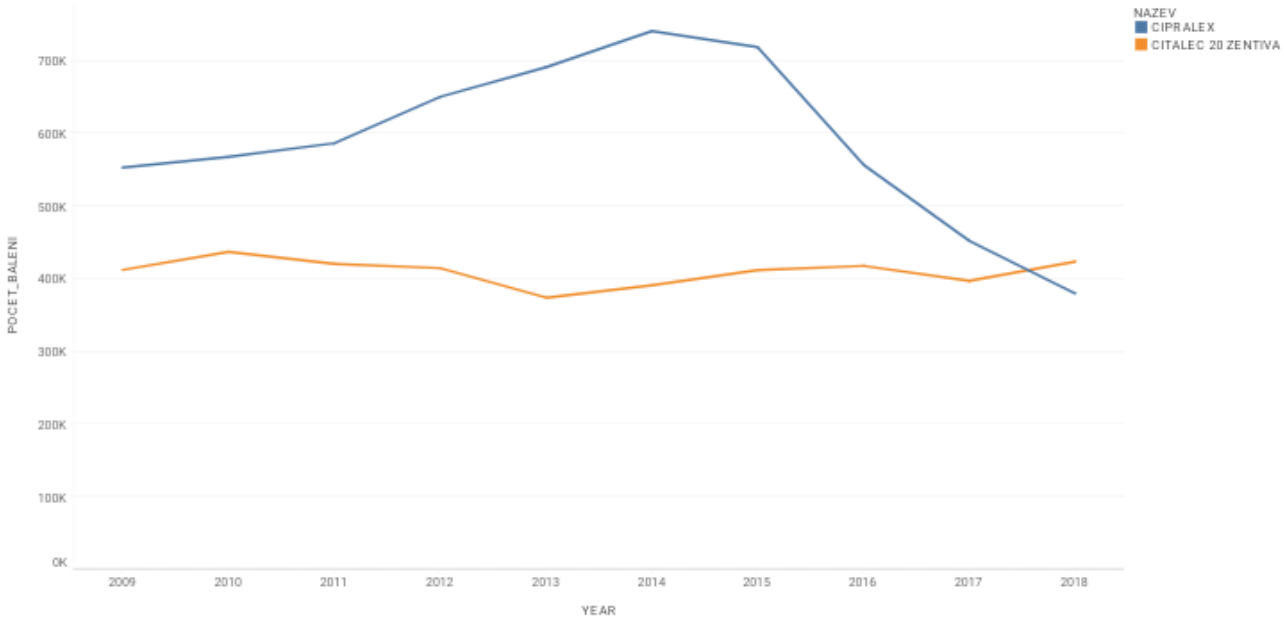


Figure 3 Cipralex and Citalec market comparison (no. of packages sold) 2009-2018.

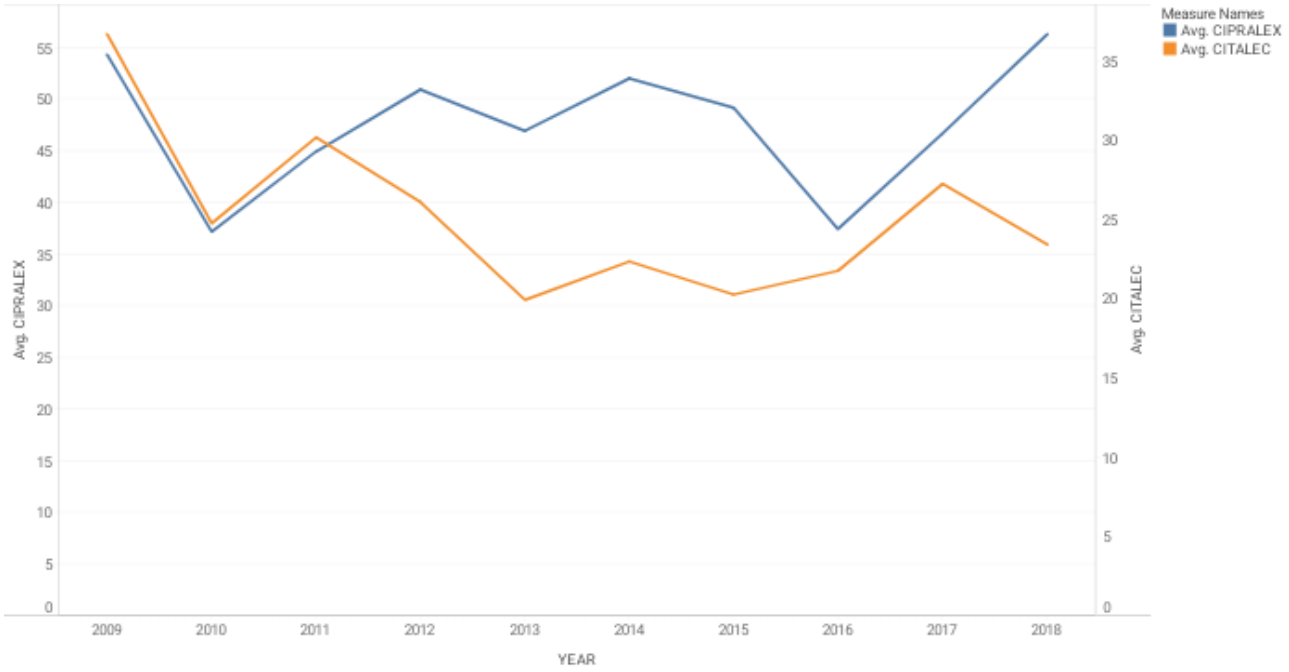


Figure 4 Cipralex and Citalec Google Trend comparison 2009-2018.

could lead to this possible usage. Firstly, the data must follow a consistent structure with clearly defined variables. Secondly, the datasets have to include the classification codes to uncover specific medicines or active ingredients. We faced significant obstacles with data synthesis during our three-month collection period across the EU member states mainly caused by poor information services with significant differences regarding open human medicine data structure and quality. Finally, we were able to demonstrate open data CI analysis with a focus on the Czech antidepressant market. Moreover, the Czech datasets provided us with the possibility of showing insights into specific medicine and company market performance thanks to rich data quality including the ATC classification, name of the producers, consumption and pricing data, as well as reliable retrospectivity. We have used the ATC classification to filter out the antidepressant class during the time period from 1991 to 2018 for the consumption trend, and then from 2009 to 2018 to provide actual market insights. Afterwards, we were able to identify key antidepressant market players and the main antidepressants used in the Czech Republic, together with consumption data (daily doses, number of packages, etc.) and finally to uncover total antidepressant consumption. Our first research question is therefore confirmed: open data and its contextual analysis bring intelligence for a specific country.

Table 5 Google Search data territorial analysis with the keyword Cipralex 2009-2018.

Czech region	***Cipralex* keyword interest**
Pardubický region	100
Region Vysočina	80
Prague	78
Ústecký region	71
Moravskoslezský region	69
Středočeský region	58
Jihomoravský region	58
Zlínský region	55
Olomoucký region	54
Královéhradecký region	50
Jihočeský region	45
Plzeňský region	43

Table 6 Google Search data territorial analysis with the keyword Citalec 2009-2018.

Czech Region	***Citalec* keyword interest**
Moravskoslezský region	100
Středočeský region	55
Jihomoravský region	54
Prague	51
Zlínský region	33

Our second research question focused on geographical aspects of open medicine data. Firstly, medicine data are strictly anonymous. None of the institutions provided datasets with geolocation. We have decided to use information-seeking behavior data and show possible geographical context to antidepressant consumption. We have used the Google Search data together with related keywords that consisted of 'antidepressant' as the general term and the specific name of the medicine. The analysis has shown that Google Search data correlates with market trends uncovered by open data analysis, but the territorial insights were not significant in this case due to the small Google Search data sample and provided only a general regional overview. However, this method would be effective in the analysis of a larger western country (e.g. United States, United Kingdom) where it is possible to work with a more significant and detailed search data sample, e.g. on a city by city level. For this reason, we consider the second research question to be confirmed.

Our future work is directed towards finding possible intelligence links between the open human medicine market data and innovation processes with the perspective of using patent data.

7. DISCUSSION

Our aim in this paper was to provide possibilities for working with open data as a tool for hard-to-get intelligence insights within the pharmaceutical sector. Not only did our results provide significant and relevant market context, but they also confirmed that open human medicine data can serve as a trend analysis information commodity for a wide range of public entities, e.g. governmental bodies for decision-making processes aimed at increasing the level of public health. Our case with antidepressants has demonstrated the trend analysis possibility within a reliable time frame. Thanks to the ATC classification system we are able to determine specific health problems within the whole population of a specific country. More importantly, we can compare countries afterwards regarding their health condition.

Our collection phase regarding the data structure and quality led us to the conclusions that open human medicine data initiatives should be considered more seriously across the EU. We have designed optimal data fields as a common base. This could determine another quality level across the whole of Europe and use open data in the most reliable way: to strengthen public health.

Still, open human medicine data in our designed structure plays an important role for intelligence studies. Based on our results, we could get insightful market information in any selected geographical area on pharmaceutical companies, medicine brands and the active ingredients of drugs together with their therapeutic, chemical and pharmacological properties.

ACKNOWLEDGEMENTS

This paper was written thanks to the long-term institutional support of research activities by the Faculty of Informatics and Statistics, University of Economics, Prague. This paper has been supported by the IGA grant "Using Open Data within Competitive Intelligence" VSE IGS F4/32/2018. We also would like to thank Dr. James Partridge for his academic support.

8. REFERENCES

Akhgar, B., Bayerl, P. S., & Sampson, F. (Eds.). (2016). *Open Source Intelligence Investigation*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-47671-1

Bernard, R., Bowsher, G., Milner, C., Boyle, P., Patel, P., & Sullivan, R. (2018). Intelligence and global health: assessing the role of open source and social media intelligence analysis in infectious disease outbreaks. *Journal of Public Health*, *26*(5), 509–514.

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., & Mandl, K. D. (2008). Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. *PLoS Medicine*, *5*(7), e151.

Cantor, M. N., Chandras, R., & Pulgarin, C. (2018). FACETS: using open data to measure community social determinants of health. *Journal of the American Medical Informatics Association : JAMIA*, *25*(4), 419–422.

Cerny, J., Potancok, M., & Molnar, Z. (2018). Open human medicine data in European Union. In Oskrdal V., Doucek P., & Chroust G. (Eds.), *IDIMT-2018 : strategic modeling in management, economy and society : 26th Interdisciplinary Information Management Talks, Sept. 5 -7, 2018, Kuntná Hora, Czech*

Republic (p. 509). Kutná Hora: University of Economics, Prague, Czech Republic.

Cook, S., Conrad, C., Fowlkes, A. L., & Mohebbi, M. H. (2011). Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *PLoS ONE*, *6*(8), e23610.

Datig, I., & Whiting, P. (2018). Telling your library story: tableau public for data visualization. *Library Hi Tech News*, *35*(4), 6–8.

European Commison. (2018). Building a European data economy.

European Commison. (2019). About | Open Data Portal. Retrieved from https://data.europa.eu/euodp/en/about

Farber, G. K. (2017). Can data repositories help find effective treatments for complex diseases? *Progress in Neurobiology*, *152*, 200–212.

Few, S. (2012). *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Analytics Press. Retrieved from https://books.google.cz/books?id=1xiHLgEACAAJ

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

Grèzes, V. (2015). The definition of competitive intelligence needs through a synthesis model. *Journal of Intelligence Studies in Business*, 5(1), 40–56.

Herring, J. P. (1999). Key intelligence topics: A process to identify and define intelligence needs. *Competitive Intelligence Review*, *10*(2), 4–14.

Hu, J., Perer, A., & Wang, F. (2016). Data Driven Analytics for Personalized Healthcare BT - Healthcare Information Management Systems: Cases, Strategies, and Solutions. In C. A. Weaver, M. J. Ball, G. R. Kim, & J. M. Kiel (Eds.), *Healthcare Information Management Systems* (pp. 529–554). Cham: Springer International Publishing.

Hughes, S. F. (2017). A new model for identifying emerging technologies. *Journal of Intelligence Studies in Business*, 7(1), 79–86.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, *29*(4), 258–268.

Kostkova, P., Brewer, H., de Lusignan, S., Fottrell, E., Goldacre, B., Hart, G., … Tooke, J. (2016). Who Owns the Data? Open Data for Healthcare. *Frontiers in Public Health*, *4*, 7.

Nuti, S. V, Wayda, B., Ranasinghe, I., Wang, S., Dreyer, R. P., Chen, S. I., & Murugiah, K. (2014). The Use of Google Trends in Health Care Research: A Systematic Review. *PLOS ONE*, *9*(10), e109583.

Oubrich, M. (2011). Competitive intelligence and knowledge creation - outward insights from an empirical survey. *Journal of Intelligence Studies in Business*, 1(1), 97–106.

Perer, A., & Gotz, D. (2013). Data-driven Exploration of Care Plans for Patients. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (pp. 439–444). New York, NY, USA: ACM.

Tableau. (2019). Tableau: Business Intelligence and Analytics Software. Retrieved from https://www.tableau.com

U.S. National Library of Medicine. (2019). MeSH Browser. Retrieved from https://meshb.nlm.nih.gov/search

WHO. (2018). ATC/DDD Index 2018. Retrieved May 29, 2018, from https://www.whocc.no/atc_ddd_index/

Journal of Intelligence Studies in Business
Vol. 9, No. 2 (2019) pp. 82-95
Open Access: Freely available at: https://ojs.hh.se/

The potential of business intelligence tools for expert finding

Mehdi Dadkhah[a], Mohammad Lagzian[a]*, Fariborz Rahimnia[a] and Khalil Kimiafar[b]

[a]*Department of Management, Faculty of Economics and Administrative Sciences, Ferdowsi University of Mashhad, Iran*
[b]*Department of Medical Records and Health Information Technology, School of Paramedical Sciences, Mashhad University of Medical Sciences, Mashhad, Iran*

Corresponding author (*): m-lagzian@um.ac.ir

Received 30 July 2019 Accepted 28 October 2019

ABSTRACT Finding the right experts for data gathering through interview serves as a key for particular research works. However, most expert finding methods in the literature require great deals of technical knowledge, making them somewhat impracticable for business researchers without deep technical knowledge. Accordingly, there is a need for an expert finding solution for researchers without a deep technical background. As business researchers may have knowledge about business intelligence and its tools, the use of business intelligence tools can be used to solve such issue. The present paper discusses the process of using business intelligence tools to find potential experts for example topics. Subsequently, based on a literature review, criteria are presented for distinguishing different experts. Finally, the analytic hierarchy process is discussed for assigning weights to both selection criteria and potential experts. The audience of this paper is researchers who are familiar with business intelligence tools or would like to learn how to work with them.

KEYWORDS Business intelligence, business intelligence tools, expert selection, expert selection criteria, participant selection

1. INTRODUCTION

In social science, qualitative methods are popular for conducting research. In the qualitative research methods, interviews with participants are one data collection instrument (Louise Barriball & While, 1994). Accordingly, different strategies are presented for selecting potential participants. In some cases, unavailability of participants for face-to-face interviews or other difficulties led researchers to utilize computers as a research instrument (Girvan & Savage, 2013; Markham, 2004). A review of the literature on research methodologies shows that, unlike quantitative research, qualitative research tends to select participants purposively (Flick, 2008; Marshall, 1996) based on specific criteria. In such studies, the researcher decides, based on the specific criteria, who to consider as a participant for the research (Flick, 2008; Marshall, 1996). There are a number of strategies for purposive sampling in qualitative research (Palinkas *et al.*, 2015). As described by Palinkas *et al.*, such strategies can be grouped into three major categories: (1) the strategies emphasizing similarity, (2) the strategies emphasizing variation, (3) and the strategies with no specific emphasis (Palinkas *et al.*, 2015; Patton, 2002). Despite the apparently extensive research on purposive sampling in qualitative research, it is not always an easy task to accomplish. It is not

always easy to find participants for a research plan where the required data shall be obtained from people with professional knowledge (i.e. experts). The situation becomes even more critical when such expert experience falls within multiple contexts, with only few experts in each context, or when such experts are in multiple locations (for example, country, university, or organization) making it impossible for the researcher to become aware of all of them. Even though snowball sampling can be a good alternative for such conditions, finding participants within a reasonably short period of time is also an issue. Finding expert participants for a qualitative research may be difficult in some cases. This problem is not limited to some cases in qualitative research: there are studies that discuss this issue without considering this domain (Gretsch, Mandl, & Hense, 2011; Ru, Xu, & Guo, 2007; Serdyukov & Hiemstra, 2008). So, finding the right expert can be a challenging task. In such situation, using a machine-made method for finding experts can be helpful.

The present research aims to show how to use business intelligence (BI) tools and the analytic hierarchy process (AHP) to find experts. The target audience of the current study is researchers who are interested in BI or have knowledge in this regard. For example, business students can learn to work with the tools used in this paper as they may learn BI in university or at workshops, Section 2 gives a brief overview of some available methods. Section 3 describes the process of using BI tools to find experts and presents a discussion on its results. The final conclusions are drawn in Section 4.

2. BRIEF OVERVIEW OF EXPERT FINDING RESEARCH

Researchers have presented various methods to find experts. Deng *et al.* presented three models for finding experts by using DBLP bibliography and Google Scholar services (Deng, King, & Lyu, 2008). Naeem *et al.* utilized data mining for the same purpose (Naeem, Khan, & Afzal, 2013). Kardan *et al.* presented and discussed a model for expert selection in social networks (Kardan, Omidvar, & Farahmandnia, 2011). Other research focuses on finding experts in social networks or community question answering websites (Bozzon, Brambilla, Ceri, Silvestri, & Vesci, 2013; Kao, Liu, & Wang, 2010; Kardan et al., 2011; Riahi, Zolaktaf, Shafiei, & Milios, 2012; Zhang, Tang, & Li, 2007; Zhao, Zhang, He, & Ng, 2014). Wang *et al.* proposed an algorithm, called *ExpertRank*, that identifys and evaluates experts based on both documentation and an individual's authority in his or her knowledge community. This algorithm is a modification of the *PageRank* algorithm to evaluate an individual's authority (Wang, Jiao, Abrahams, Fan, & Zhang, 2013). Demartini used Wikipedia as the knowledge source to find experts in topics. He used *WordNet* and *Yago* to improve retrieval effectiveness (Demartini, 2007). Zhan *et al.* employed probabilistic latent semantic analysis to propose a mixture model for expert finding. Semantic themes will be identified by such mixture models between terms and documents. Then by using these themes, their method finds relevant experts based on the query (Zhang, Tang, Liu, & Li, 2008). Yang *et al.* proposed an expert finding system by analyzing an individual's journal papers. They state that journal publication can be used to find the expertise of a researcher (Yang, Chen, Lee, & Ho, 2008). Lin *et al.* in a survey discussed methods and models that focus on expert findings and show the current status of research in this regard (Lin, Hong, Wang, & Li, 2017). Boeva *et al.* proposed a data driven expert finding technique. Their technique also weighs experts based on their expertise (Boeva, Angelova, & Tsiporkova, 2017). Further search into the literature would highlight other technical methods for expert finding.

Even though these are valuable and interesting, such methods are only useful for researchers with advanced technical knowledge. Other researchers without deep technical knowledge may not be able to take advantage of such techniques, unless the technical methods are translated into convenient tools for social science researchers. There are some easy-to-use expert finding methods in the literature. On its user interface, Scopus provides an interested option for analyzing search results (Beatty, 2015), offering an easy-to-use method for non-technical researchers who are looking for particular experts. This method can be used for expert selection. Schuemie and Kors developed a web-based tool entitled *Jane* (http://jane.biosemantics.org/index.php) which can be used for expert finding. *Jane* uses PubMed as the source of data and presents result by using the Lucene MoreLikeThis algorithm and k-nearest neighbor approach (Schuemie & Kors, 2008). Cifariello *et al.* developed a semantic search engine entitled

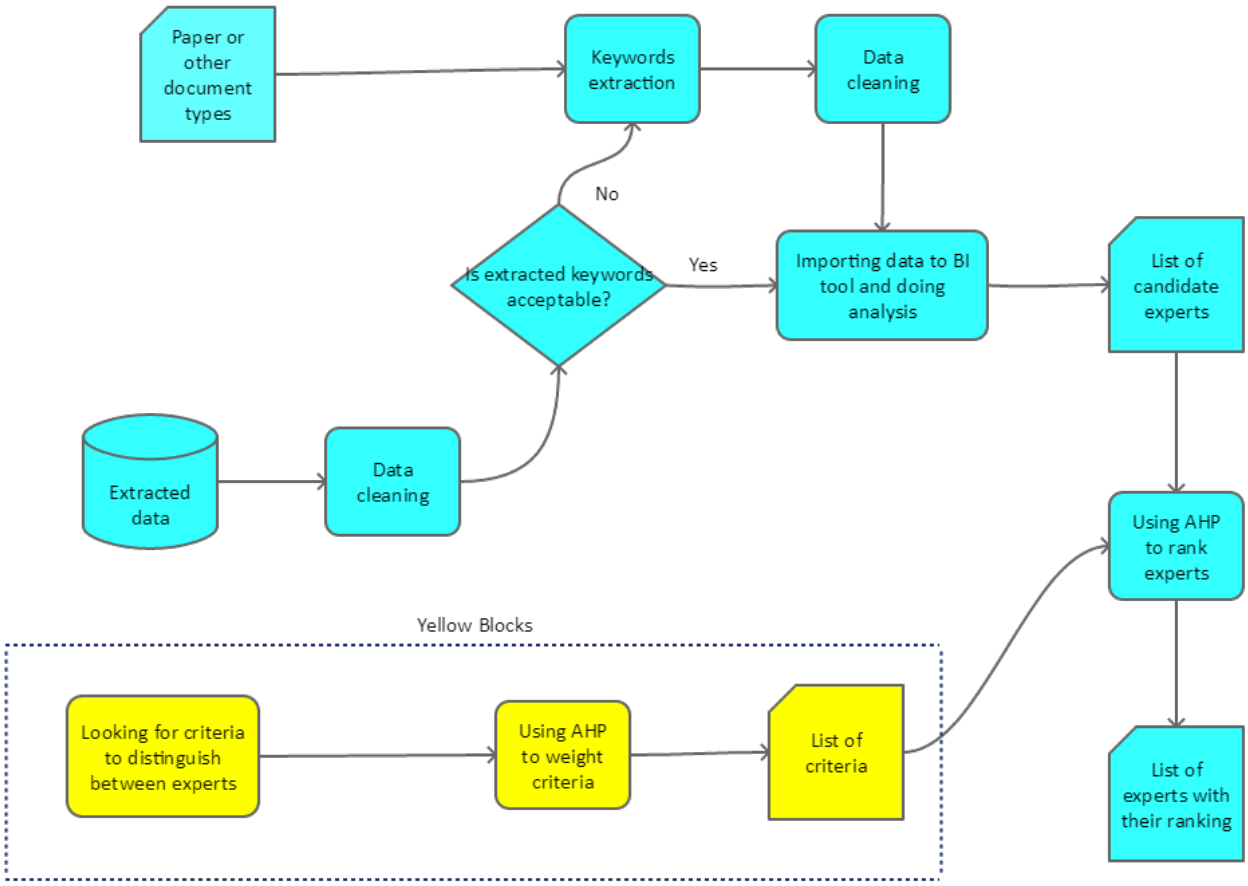


Figure 1 Schematic presentation of the proposed method for expert selection.

Wiser that finds experts. It models each author's expertise with a graph by using Wikipedia. Experts are identified through co-occurrence of searched keywords in their publications and this graph. Wiser has an online graphical-based version (https://wiser1.sobigdata.d4science.org/search) which works based on University of Pisa publications (Cifariello, Ferragina, & Ponza, 2019). These tools help researchers to find experts. However, when a user with no advanced technical knowledge aims to analyses his or her own data or data related to other academic sources, this method is not helpful. It should be noted that is possible to the adapt proposed method in literature to be used for different data sources, but technical knowledge in this regard is required. BI tools are especially useful for business students to find experts. This study focusses on a process that helps researchers to find experts by utilizing BI tools. The process in this paper can be used by individual who are familiar with BI to find experts. This paper does not present a new method, it shows the capability of existing BI tools to be used for expert finding.

3. PROCESS OF FINDING EXPERTS USING BI TOOLS

Today, organizations are encountering large sets of data that cannot be used without BI. In order to make better decisions, organizations utilize BI to create knowledge out of their data (Chaudhuri, Dayal, & Narasayya, 2011). A BI solution follows a BI architecture. Generally, companies store different data of different sources. However, before a BI solution can be successfully implemented, the entire set of such data must be integrated to a data warehouse by using a special process called ETL (extract, transform and load). Given the inefficiency of executing queries on an entire set of data in an organization, it is necessary to extract related data before proceeding to executing such a query. Once an integrated data warehouse is developed, different servers can efficiently access the data in the warehouse through front-end applications. Such an application can be used by particular decision-makers depending on their roles in the organization (Chaudhuri *et al.*, 2011; Negash, 2004; Sherman, 2014). Details of BI are out of scope of the present work, where only the BI tool is used, rather than a full BI implementation. A BI tool is a vendor's software that is used to develop BI applications or styles (e.g. dashboards or scorecards) (Sherman, 2014).

Table 1 The data extracted from Scopus by searching the term "Internet of Things".

Author Ids	Title	Year	Source title	Author Keywords
Records of data				
14018777000; 27867946500; 57202208939; 57202211443; 38461465700;	Multidimensional wavelet neuron for pattern recognition tasks in the internet of things applications	2019	Advances in Intelligent Systems and Computing	Classification; Internet of things; Machine learning; Multidimensional wavelet neuron; Online learning; Pattern recognition
57202334348;	FAN: Framework for authentication of nodes in mobile ad hoc environment of internet-of-things	2019	Advances in Intelligent Systems and Computing	Access control; Internet-of-Things; Mobile ad hoc network; Secure permission; Security; Ubiquitous
57203555315; 56238720400;	Study and design of smart embedded system for smart city using internet of things	2019	Lecture Notes in Electrical Engineering	Electronic devices; Internet of Things (IoT); Smart city
Other records of data				

Table 2 The cleaned data for the analysis in this study.

Author Ids	Title	Year	Source title	Author Keywords
Records of data				
14018777000	Multidimensional wavelet neuron for pattern recognition tasks in the internet of things applications	2019	Advances in Intelligent Systems and Computing	Classification; Internet of things; Machine learning; Multidimensional wavelet neuron; Online learning; Pattern recognition
27867946500	Multidimensional wavelet neuron for pattern recognition tasks in the internet of things applications	2019	Advances in Intelligent Systems and Computing	Classification; Internet of things; Machine learning; Multidimensional wavelet neuron; Online learning; Pattern recognition
57202208939	Multidimensional wavelet neuron for pattern recognition tasks in the internet of things applications	2019	Advances in Intelligent Systems and Computing	Classification; Internet of things; Machine learning; Multidimensional wavelet neuron; Online learning; Pattern recognition
57202211443	Multidimensional wavelet neuron for pattern recognition tasks in the internet of things applications	2019	Advances in Intelligent Systems and Computing	Classification; Internet of things; Machine learning; Multidimensional wavelet neuron; Online learning; Pattern recognition
38461465700	Multidimensional wavelet neuron for pattern recognition tasks in the internet of things applications	2019	Advances in Intelligent Systems and Computing	Classification; Internet of things; Machine learning; Multidimensional wavelet neuron; Online learning; Pattern recognition
57202334348	FAN: Framework for authentication of nodes in mobile ad hoc environment of internet-of-things	2019	Advances in Intelligent Systems and Computing	Access control; Internet-of-Things; Mobile ad hoc network; Secure permission; Security; Ubiquitous
57203555315	Study and design of smart embedded system for smart city using internet of things	2019	Lecture Notes in Electrical Engineering	Electronic devices; Internet of Things (IoT); Smart city
56238720400	Study and design of smart embedded system for smart city using internet of things	2019	Lecture Notes in Electrical Engineering	Electronic devices; Internet of Things (IoT); Smart city
Other records of data				

Partially inspired by the general BI solution, and its uses for academic research introduced by Chaudhuri et al. (2011), Sherman (2014) and Dadkhah and Lagzian (2018) the process of experts finding is schematically presented in Figure 1. Similar to the work by Boeva *et al.*, the process herein uses a keyword-based search to identify experts (Boeva, Angelova, & Tsiporkova, 2017). A basic requirement of a BI process is data. The data may come from different sources. In the field of research, such data may be collected from academic databases such as Scopus or Google Scholar, academic papers, un-published documents, or reports. For the most part, the academic databases provide the user with an option to extract relevant data based on various criteria. For example, upon searching Scopus for the term "Internet of Things", one can extract the *titles*, *authors' names*, *keywords*, and/or *names of the journals* corresponding to the search, resulting in a file of a particular format. Figure 1 highlights such data as "extracted data". When it comes to possibly large offline documents on a local disk, there is a need for methods to either automatically extract such data and print that into a file or do the same manually. Various methods have been proposed for keyword extraction in the literature (MATSUO & ISHIZUKA, 2004; Merrouni, Frikh, & Ouhbi, 2016; Rose, Engel, Cramer, & Cowley, 2010). In such processes, keywords play a fundamental role. The present work is focused on two features in each document: the author's name and keywords. Table 1 shows a summary of the data extracted from Scopus by searching the term "Internet of Things", as an example. This search was limited to 2000 records by the authors (search date: 7 September 2018). Accordingly, the following features were included in the data: Author Id, Title, Year, Source title, Author Keywords.

Upon extracting the relevant data, one should check for possible inconsistencies, errors or related issues. For example, there may be duplicate records to be cleaned up or inconsistencies to be addressed by reformatting the data. The data cleanup stage is critical for the successful accomplishment of the entire process. In the present work, an easy-to-use freeware called OpenRefine was used to clean up the data ("OpenRefine," 2018) (Verborgh & De Wilde, 2013). After the cleanup stage, one should evaluate the acceptability of the extracted keywords. If the keywords were found to be unacceptable, automatic keyword extraction methods can be applied to extract other keywords. Table 2 shows the extracted data following the cleanup stage. As suggested by the designation, *Author Id*s indicate the authors' names and help classify keywords by authors. Accordingly, a single *Author Id* was presented per row. Also, correction may be necessary for spelling multiplicity in the source title. The records lacking an *Author Id*, with the corresponding field left blank, were deleted in this study. In Table 2, each row refers to a particular author and provides details of paper title, year of publication, place of publication, and keywords.

At this stage, the dataset is ready for analysis. This paper deals only with the BI tool rather than a full BI implementation. There are different BI tools with different features, and their associated costs vary from free to paid. BI tools provide different features including dashboards and reporting capability. Dashboards provide graphical elements for data visualization. Reporting capability lets the user use the information element (Bernardino & Tereso, 2013). Both reporting and dashboard elements can be used to find relevant experts. In this paper, a trial licensed version of DBxtra (https://dbxtra.com) was used as we had access to it, and it provided a drag-and-drop option. The documentation of this tool provides a good source for operating the software (DBxtra, 2018). Utilizing the software, a constraint was set to consider only records for which at least two features were available: author's name and keywords. Then the authors were filtered based on keywords to find relevant experts. This is why the present method was said to be based on keywords. For example, we filtered authors by selecting the keywords "energy", "sensor" and "IoT", then the software listed the authors who published papers contained these terms as keywords. In DBxtra, a dashboard is designed using a list box, two combo boxes, a chart, and a pivot table. For the example considered in this research, the list box contained the *Author Keywords* values. Accordingly, a list of relevant experts could be obtained by applying a filter on this list. As shown in Figure 2, a filter was designed to extract the list of authors who had used the terms "energy", "sensor" and "IoT" as keyword.

The two combo boxes could filter the data by year and place of publication, with the chart indicating the count of candidate experts.

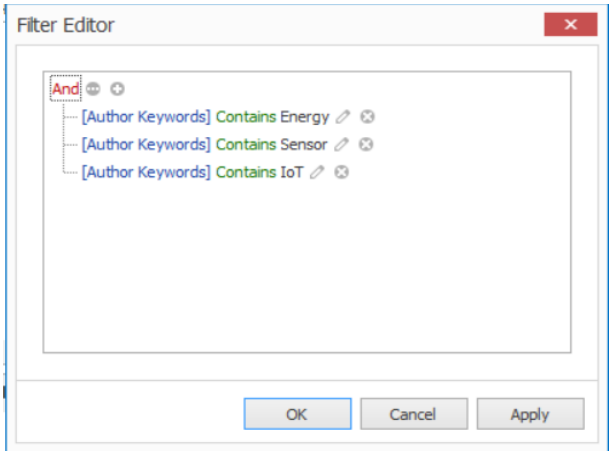


Figure 2 The filter applied on the list box to extract the list of authors who had used the terms "energy", "sensor" and "IoT" as keyword.

Figure 3 shows the dashboard with the experts who had paper(s) containing the following keywords: "energy", "sensor" and "IoT". Accordingly, a list of 66 experts with expertise related to sensors and energy in the IoT domain was obtained. There are different dashboard elements that researchers can refer to in order to document and understand their tools. By using such elements, there is the possibility to visualize data and do relevant analyses, then find experts. When the data is clean, the availability of working with BI tools and their elements plays an important role in finding relevant experts from data. Based on their needs, researchers should decide which elements are helpful for their analysis add them to their dashboard. Also, each element needs to be configured. For example, the chart in Figure 3 is configure to count the number of Author IDs in the data. Generally, it counts a distinct value of Author IDs in all data. The combo box is configured to include data related to the keywords. When a filter is applied on this combo box, the chart counts only the Author IDs that are accessible through this filter. We do not discuss more about the capabilities of each BI tool and their related elements, as there is good documentation in this regard.

3.1 Ranking experts based on the research topic

Once one is finished identifying the relevant experts, it is possible to evaluate the suitability of such potential experts for the research. The BI tool provides potential experts and next the researchers should confirm result. They should evaluate each potential expert to understand if the person is a suitable expert. As an example, one may need only 10 experts. If the BI tool provided 66 experts (Figure 3), one must select the 10 most suitable experts. For this purpose, beginning with an attempt to distinguish between experts based on some general criteria, one should remember that specific research exists with additional features for the purpose. In this study, relevant features were

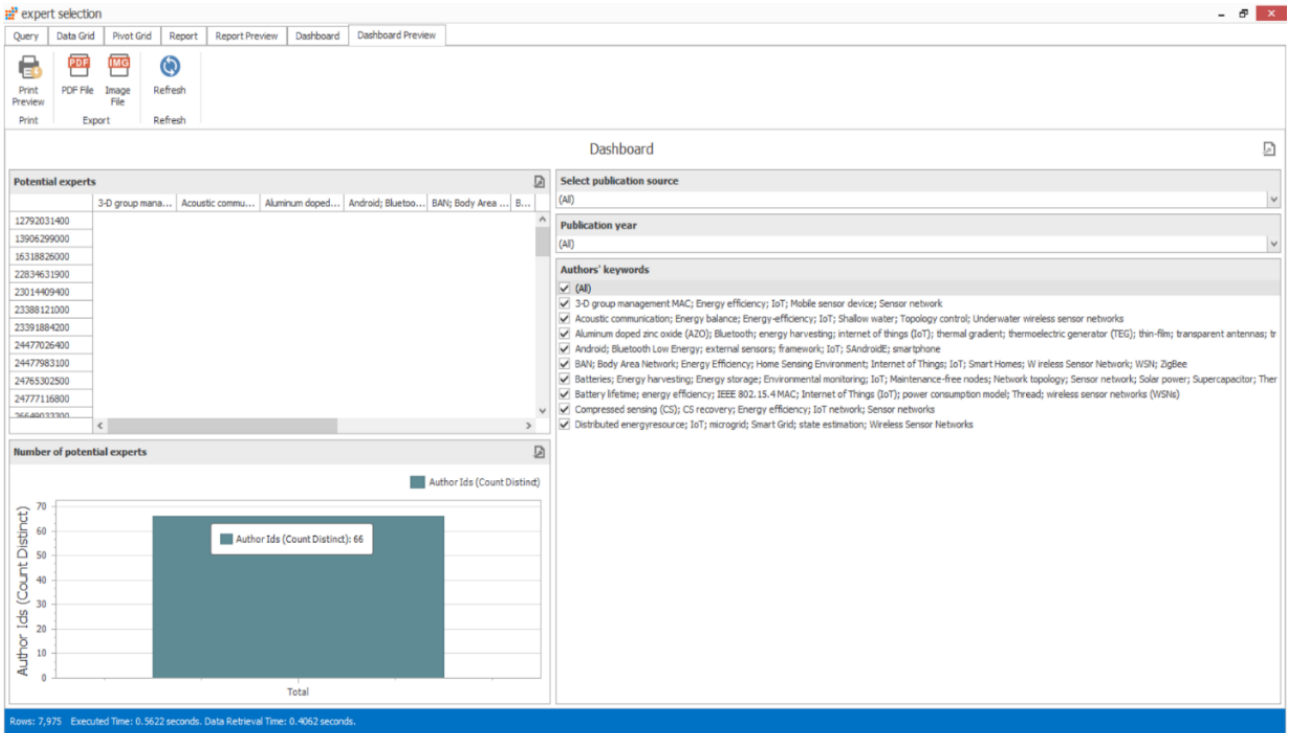


Figure 3 The dashboard designed for finding experts who had published paper(s) containing the following keywords: "energy", "sensor" and "IoT".

identified by looking into the literature. Accordingly, papers presenting criteria for expert selection were identified (Afzal, Kulathuramaiyer, & Maurer, 2008; Afzal & Maurer, 2011; Benner, Tanner, & Chesla, 1992; Boeva et al., 2017; Cameron, Aleman-Meza, Decker, & Arpinar, 2007; Hirsch, 2005; Naeem et al., 2013; Quatrini Carvalho Passos Guimarães, Pena, Lopes, Lopes, & Bottura Leite de Barros, 2016); (Academia Europaea as cited in Naeem *et al*., 2013; Pakistan Academy of Sciences as cited in Naeem *et al*., 2013; Fehring as cited in Quatrini Carvalho Passos Guimarães, Pena, Lopes, Lopes, & Bottura Leite de Barros, 2016). As some of these papers were subject-oriented, respective criteria were generalized and used as a feature for expert identification and ranking (Table 3). Researchers may need to define new criteria based on their research.

In the next step, an analytic hierarchy process (AHP) can be used to assign weights to the features to facilitate the process of decision-making for expert selection. AHP refers to a pairwise comparison method for weighting a pool of alternatives, so as to select an alternative based on particular criteria. Using multilevel hierarchic structures, an AHP involves alternatives, criteria, and a goal. It has been widely used in business- and government-led applications (Saaty, 1977, 1990, 2013). In this paper, AHP is utilized to rank a set of candidate experts based on particular criteria extracted from the literature, for the purpose of final expert selection. AHP arranges the decision criteria into a hierarchical structure. In this stage, the scale shown in Table 4 can be used as a foundation to design a questionnaire for pairwise comparison (Saaty, 1977, 1990, 2013).

Table 3 The features used for selecting and ranking the experts (adapted from the references cited in the text).

No.	Feature	Description
1	Projects	To distinguish experts participating in a particular project(s).
2	Awards	To distinguish experts who have achieved a particular award(s)
3	Honorarium	To distinguish experts who have contributed into a particular domain(s).
4	Affiliations	To distinguish experts with a particular affiliation(s), taking the affiliation as a measure of proficiency in a particular domain(s).
5	Request for Comments (RFC)	To distinguish experts who were frequently requested for comments, taking RFC as a measure of experimental skills in a particular domain(s).
6	Supervision	To distinguish experts who are active in the field of academic supervision of students.
7	Collaboration	To distinguish experts who have collaborated with others at international level.
8	Relevance	To distinguish experts who are actually relevant to the considered research.
9	Keynote Speaker	To distinguish experts who have been a keynote speaker in a conferences or other societies.
10	Reviewer	To distinguish experts with the required deals of skill and expertise to serve as a reviewer for a journal or conference.
11	Protocol Design	To distinguish experts with the required deals of skill and knowledge to design protocol standard(s).
12	Distinctions	To distinguish outstanding experts, in comparison to peers.
13	Citation number	To distinguish experts with a particular number of received citations.
14	Publication number	To distinguish experts with a particular number of publications.
15	Co-author network	To distinguish experts who have worked with a particular number of co-authors.
16	Academic degree	To distinguish experts with a particular academic degree.
17	Gender	To distinguish experts of a specific gender.
18	Experience duration	To distinguish experts with a particular number of years of contribution into the considered domain.
19	Extent of citations in given domain	To distinguish experts based on the number of received citations in a particular domain: $Extent\ of\ Citation = \frac{total\ number\ of\ received\ citation\ in\ a\ topic\ by\ candidate\ expert}{total\ number\ of\ received\ citation\ in\ a\ topic}$
20	Impact factor of publication journals	To distinguish experts who had papers published in journals of particular impact factor(s).
21	H-Index	To distinguish experts based on the metric proposed by Hirsch. This metric indicates the j number of papers that received j or higher number of citations.
22	Researcher profile	To distinguish experts based on their profile in terms of relevant skills, keywords, and topics of interest.

Table 4 Scales for comparing alternative experts (Saaty, 1977, 1990, 2013).

Numeric scale	Meaning
1	The two alternatives are equally important.
3	An alternative is moderately more important than another.
5	An alternative is essentially more important than another.
7	An alternative is strongly more important than another.
9	An alternative is extremely more important than another.
2, 4, 6, 8	Intermediate values between the above milestones.

The yellow blocks in Figure 1 show the corresponding steps through the whole process. If there is uncertainty in decision making, fuzzy AHP can be used. It uses fuzzy numbers as the numerical scales (Özdağoğlu and Özdağoğlu, 2007; Wang and Chin, 2011; Ramík and Korviny, 2010). the value of features in Table 3 should be gathered or calculate manually for each candidate expert, but it is possible to use a programming language to automate some tasks.

In order to implement AHP in this study, the experts list and the features described in Table 3 were taken as the alternatives and criteria, respectively (Figure 4). Then, two pairwise comparison questionnaires can be designed for the considered criteria and experts. The questionnaires should be presented to a number of university professors and researchers in the field of research. The data extracted from the questionnaires can be analyzed using different tools such *Super Decisions*, a tool for multi-criteria decision making (SuperDecsion, 2018), and the results should be used to assign weights to the criteria and experts. Researchers usually need to select and rank such criteria for their research activities, as may be necessary depending on the specific research question(s). Figure 4 shows the hierarchy of the AHP model developed for expert selection. In this study, weights are calculated for each criterion and the BI tool provides a list of potential experts as alternatives for this model (Figure 4). In a final step, the experts were ranked based on the criteria.

In AHP, the consistency ratio shall be equal to or smaller than 0.1; otherwise the result of pairwise comparison may be unreliable (Saaty, 1977, 1990, 2013). Indeed, the consistency ratio increases when increasing the number of elements in a comparison (Benítez *et al.*, 2011). Accordingly, the use of 22 criteria or an expert list with many candidates in the developed AHP model may lead to a consistency ratio exceeding 0.1, indicating unreliable results. However, researchers could choose to select only a subset of the 22 criteria, depending on the scope of their research, or narrow their queries to find a smaller number of experts. Other methods have also been proposed for addressing the problem of inconsistency in AHP (Benítez *et al.*, 2011; Benítez *et al.*, 2012). The value of features shown in Table 3 should be determined manually by researchers, however it is possible to gather values for some of the features automatically. For example, h-index and total citation count, number of

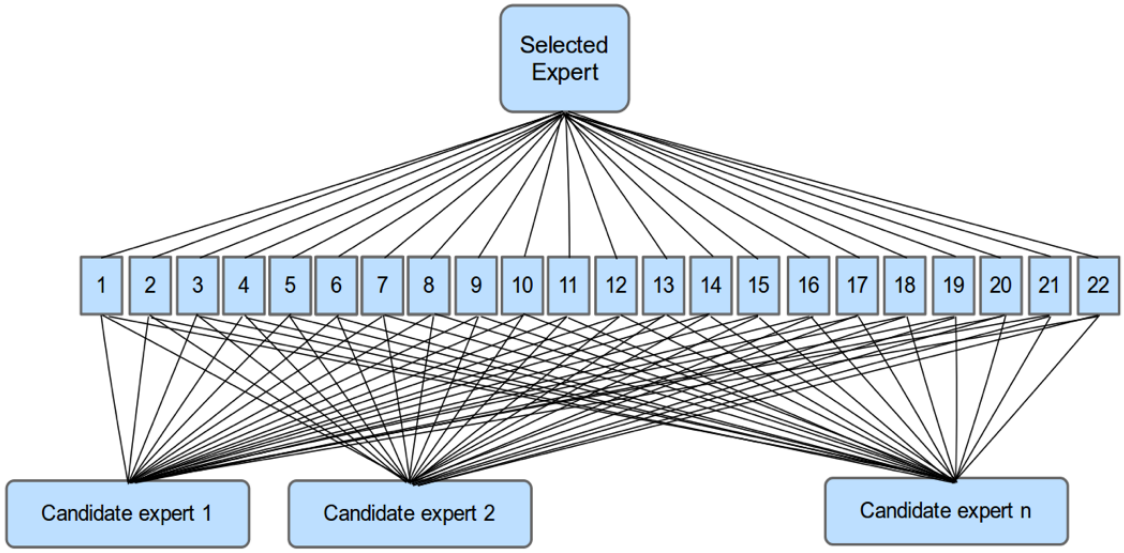


Figure 4 The AHP model developed for expert selection.

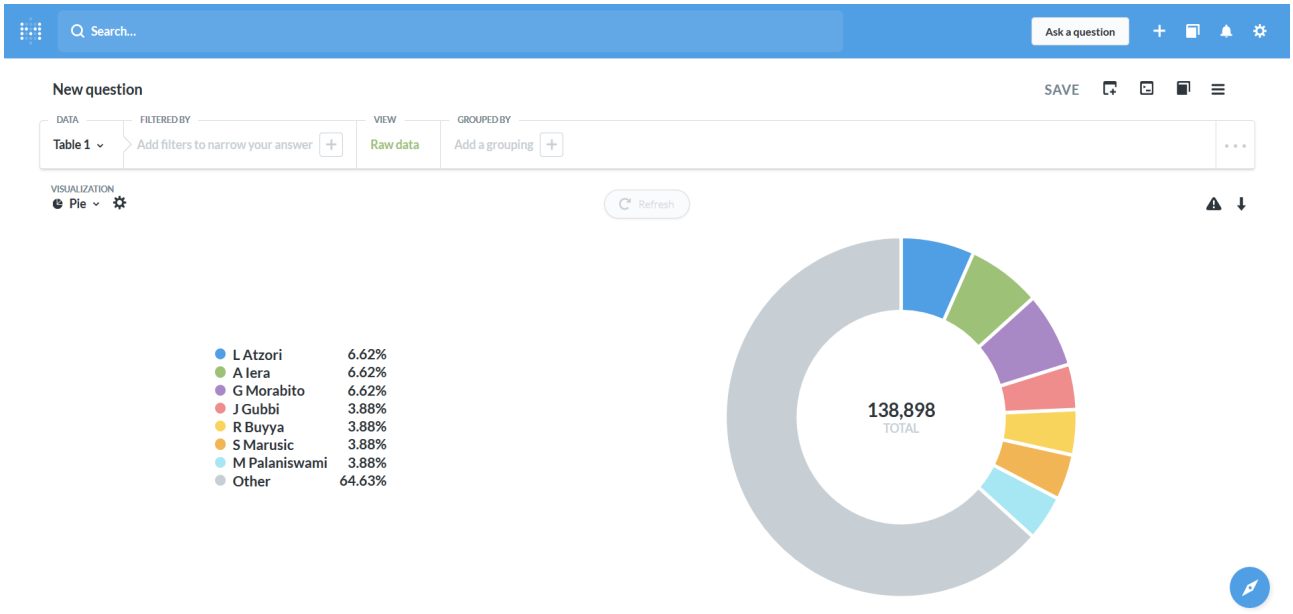


Figure 5 Identified experts using Metabase.

publication and co-authors can be gathered through Scopus.

By using BI tools, it is possible for researchers to do more advanced analysis on their data. For example, by extracting data from Scopus by searching the term "Internet of Things", it is possible to find experts with different conditions such as:

> *Experts who published a paper about the Internet of Things AND started their publication in this topic at least 5 years ago AND have a total citation count on this topic above 1200 AND have the article type "Journal Paper" AND are affiliated to a specific country AND published by a specific publisher AND published in a top information system journal.*

It is possible to add four columns including journal impact factor, author h-index, publication number, and total number of co-authors to the extracted data from Scopus. Here we attach new data to the extracted data from Scopus. This data is the value of the four features discussed in Table 3. Now, the previous query could be more advanced as:

> *Experts who published a paper about the Internet of Things AND started their publication in this topic at least 5 years ago AND have citation counts on this topic higher than 1200 AND their article type is a Journal Paper AND are affiliated to a specific country AND published by a specific publisher AND published in a top information system journal AND published in a journal with an IF higher than 1 AND with a total number of published papers higher than 10 AND author's h-index is higher than 5 AND total number of co-author is higher than 12*

This process can be done through other tools and data sources. To evaluate this expert finding process, researchers used two other tools and tried to find potential experts who are familiar with both the internet of things (IoT) and patient monitoring. The researchers are interested in experts who received at least 700 citations on a publication in this topic and published it at least five years ago. They used Publish or Perish (Publish or Perish, 2018) to extract data from Google Scholar and Metabase (Metabase, 2018) to analyses the data (search date: 10 November 2018). As Publish or Perish does not provide keywords for each paper, there are two option to find keywords: 1) use methods for extracting keywords from papers, 2) narrow the search by defining all keywords then analyzing the result instead of doing a broad search and then limiting result by keywords. Figure 5 shows the output of the analysis in Metabase. Based on this analysis, the researchers found 15 potential experts. In the extracted data from Google scholar via Publish or Perish, there are other features including Source Title, Publisher, Article URL, Cites Per Year, Author Count, and Title of Papers. This means that it is possible to use these features to do more advanced searches to find potential experts

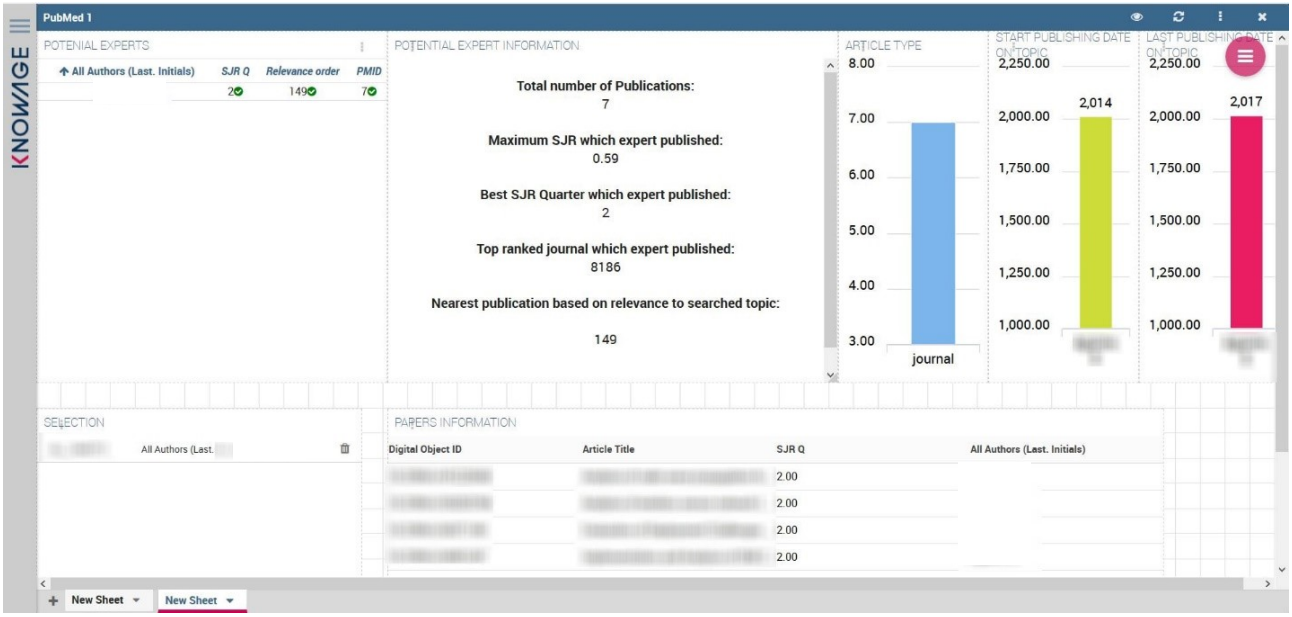


Figure 6 A dashboard in Knowage for finding experts. In this dashboard, an expert has been selected and the information is shown.

from this data. After finding expert via the BI tool, now we can manually review experts and use the Table 3 criteria to confirm experts with regard to our research.

Experts can also be found via PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) as the source of data, and Knowage (https://www.knowage-suite.com/site/home/) as the business intelligence tool. We searched for *"wireless sensor network"* (search date: 24 July 2019) and download all 769 result as the XML file. By using PubMed2XL (available from http://blog.humaneguitarist.org/projects/pubmed2xl/), the XML file was converted to an Excel spread sheet (Isaak, 2016). By using OpenRefine, the data was clean. As with *Jane*, it is possible to find potential experts based on the relevance of keywords. In PubMed, data is sorted according to its relevance to the search term, then downloaded. By having relevance of data to searched terms, it is possible to find experts based on relevance. By doing this, it is concluded that from the top 20 identified potential experts in Knowage, 15 of them were also in the list of retrieved experts from *Jane*. The difference was their rank compare to the *Jane* result. It is possible to get a list of potential experts who are familiar with wireless sensor networks by using the extracted data from PubMed. For example, we can find all individuals who have at least four publications about wireless sensor networks and at least one publication in the top 300 results, based on relevance. It is possible to do a more advance query to find individuals who have at least four publications about wireless sensor networks and at least one publication in the top 300 results based on relevance and at least one publication published in a journal in the first two quarters of the Scimago journal ranking (SJR). This query needs to merge new data with extracted data from PubMed. The *SJR* data can be retrieved from the SCiMago journal ranking website (available from https://www.scimagojr.com/journalrank.php). Then it is possible to merge the data together by using available tools such as OpenRefine. Figure 6 illustrates the dashboard in Knowage for finding such experts. This dashboard also has extra filters to find experts such as the first year of publication and the number of publications in a top quarter journal. It also shows some information about experts and their relevant papers.

The process discussed in this paper was also tested to find research method experts from a personal repository, and another study about knowledge management. Using this method was helpful for both this study and to simplify the expert finding task. In the earlier expert finding task, eight potential experts of the former 10 potential experts were identified. The main advantage of the process compared to most expert finding methods is that it has lower requirements for individual BI tool technical knowledge. BI tools currently support different options (for example drag and drop) to simplify the data analysis task (Smuts, Scholtz, & Calitz, 2015). By using a BI self-service tool, individuals can use BI tools with less technical knowledge (Imhoff & White, 2011).

Table 5 Comparison between Jane, Wiser and the proposed process in this paper. This table considers the currently available tools, not the techniques that are behind them. For example, Wiser can be used on different data sources, but in the currently available version, it is based only on University of Pisa publications. *Publish or Perish* is not an expert finding tool, it is an effective citation analysis tool that can be used for expert finding purposes. We recommend to import output data of Publish or Perish in BI tools for expert finding purposes.

Name*	Data source	Level of required knowledge	Capability for defining criteria by user	Visualization capability	Expert ranking
Jane	PubMed	No special knowledge, easy to use	Limited criteria can be defined based on advanced search option in the tool UI	No	Yes, automatically
Wiser	University de Pisa publications	No special knowledge, easy to use	There is no option for defining criteria in wiser UI	Yes	Yes, automatically
Publish or Perish	Web of science, Scopus, Crossref Google Scholar, and Microsoft Academic Search. It is also possible to import external data	primarily knowledge about scientific bases and citation analysis is necessary	User can define some criteria	No	It is possible to rank expert based on output values. For example, sorting based on h-index
Proposed process	Publications data from different sources such as Scopus, or Google Scholar	Primarily knowledge about data, scientific databases and data tools necessary	User can define different criteria as there is data to support such criteria	Yes, by using BI tools visualization elements	Yes, manually by using AHP and automatically by defining in BI tools

This process can be compared with two main expert finding approaches: manual expert finding by searching in scientific databases and proposed technical methods in the literature. Researchers can use scientific databases such as Google Scholar or Scopus to search for keywords and manually inspect search result to find experts. The process in this paper has other advantages including:

- In the manual inspection of result, researchers cannot consider all results and are limited in the publications that they can analyze in terms of time and effort.
- Researchers cannot execute an advance query on search result without utilizing BI tools without advanced technical knowledge.
- When data are collected from other sources, such as organizational publications or internal repositories, it is not possible to use Scopus or scientific databases to import data for analysis.
- When data come from internal repositories, they may be in different topics and domains, thus, manual inspection of such data may require significant time and effort to assess.

In comparison with proposed technical methods in the literature for expert finding, this process is easier in terms of implementation for researchers who have BI knowledge but do not have advance technical knowledge. If technical methods are the tool implemented and are publicly accessible for all researchers, they can be compared in terms of capabilities and advantages with BI tools. For that purpose, a comparison between Jane, Wiser and the process in this paper is shown in Table 5.

The process in this study may be limited to cases where data is related to the potential experts' publications. Future research can focus on using BI tools to find experts based on data gathered from social networks or community question answering websites. In this paper we only focus on the usefulness and level of required technical knowledge to evaluate this process with the proposed methods in the literature. The main goal of this study is to propose a simpler expert finding process, which provides acceptable results based on analyzing publications, not providing a comprehensive expert finding method. The

contribution of this research is a discussion on a process for finding experts by using BI tools. This paper does not propose a new tool or method, but it introduces the capability of existing BI tools for finding potential experts.

4. CONCLUSION

Given that the existing expert selection methods are usually impractical for researchers without deep technical knowledge, an expert selection process is discussed here for individuals who are familiar with BI tools. Taking advantage of BI tools, the process was found to have a large potential for expert finding. The process will be helpful in research that aims to gather data from expert participants. Here, we may need the opinions of experts and finding these experts is key.

The process in this paper requires a certain level of technical knowledge, because the method for expert finding is based on computers, which are technical in nature. The primarily knowledge about data, scientific databases and data tools is necessary for individuals who aim to use BI tools for expert finding. However, such knowledge can be obtained by participating in a workshop or reading relevant books and tutorials. This process is simpler, when we are aware of BI tools that support different options to simplify tasks, such as providing drag and drop options (Smuts, Scholtz, & Calitz, 2015). In addition, there are efforts for providing self-service BI tools which individuals can use with less technical knowledge (Imhoff & White, 2011). However, utilizing expert knowledge of programming helps researchers to collect more complete data and execute more complex queries. Also, for advanced data analysis, the knowledge of programming may be essential. Researchers, by improving their skills, could gain more benefit from this process. BI tools have the potential for data visualization and analysis, but related skills are required for such capabilities be reachable. In this paper, BI tools have been used to find an early list of potential experts from the data, then AHP helps to manually distinguish them and produce a final list of experts. Based on available data, a primary filtering of the list of many experts is done through BI tools, then by using AHP, a final list of experts is identified manually. So, queries in the BI tool may be simple, for example finding experts who have a total of more than 1000 citations. Such queries will make a limited list of potential experts, which is usable in AHP. The threshold and criteria for early filtering of experts using BI tools can be defined by consulting with experts. All thresholds in the presented cases in this paper are examples. In the actual expert finding process, consulting with experts to identify threshold and selection criteria based on available data for early filtering of experts is required. This process helps researchers to find experts for their work, even they are not experts in BI tools. However more knowledge and skills are needed for BI tools, to make them more successful in finding suitable experts.

ACKNOWLEDGEMENTS

It is our pleasure to thank DBxtra company for their support and providing a trial license of their tool for our work. We also appreciated assistance of researchers who helped us to understand their methods or answered our questions regard to this research. They are: Professor Robert Davison, City University of Hong Kong; Professor Dr. Muhammad Tanvir Afzal, Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan; Professor Dr. h.c. mult. Hermann Maurer, Computer Science, Graz University of Technology, Graz/ Austria; Dr. Muhammad Naeem, researcher in mobility, VeDeCoM, Versailles France; Prof. Steven Gordon, Technology, Operations and Information Management Division, Babson College, Babson Park, MA; and other researchers and companies who helped us.

5. REFERENCES

Afzal, M. T., Kulathuramaiyer, N., & Maurer, H. (2008). Expertise finding for an electronic journal. *proceedings of I-Know (Graz, Austria*, 436-440.

Afzal, M. T., & Maurer, H. A. (2011). Expertise Recommender System for Scientific Community. *J. UCS, 17*(11), 1529-1549.

Beatty, S. (2015). Analysis tools-Analyze thousands of search results in less than a minute Retrieved from https://blog.scopus.com/topics/analysis-tools

Benner, P., Tanner, C., & Chesla, C. (1992). From beginner to expert: gaining a differentiated clinical world in critical care nursing. *ANS Adv Nurs Sci, 14*(3), 13-28.

Bernardino, J., & Tereso, M. (2013). *Business Intelligence Tools*, Dordrecht.

Boeva, V., Angelova, M., & Tsiporkova, E. (2017). *Data-driven Techniques for Expert Finding.*

Paper presented at the International Conference on Agents and Artificial Intelligence.

Bozzon, A., Brambilla, M., Ceri, S., Silvestri, M., & Vesci, G. (2013). *Choosing the right crowd: expert finding in social networks*. Paper presented at the Proceedings of the 16th International Conference on Extending Database Technology, Genoa, Italy.

Cameron, D. H. L., Aleman-Meza, B., Decker, S., & Arpinar, I. B. (2007). *SEMEF: A taxonomy-based discovery of experts, expertise and collaboration networks.* University of Georgia.

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Commun. ACM, 54*(8), 88-98. doi:10.1145/1978542.1978562

DBxtra. (2018). DBxtra Online Documentation. Retrieved from www.dbxtra.com/documentation/

Demartini, G. (2007). *Finding experts using wikipedia.* Paper presented at the Proceedings of the 2nd International Conference on Finding Experts on the Web with Semantics-Volume 290.

Deng, H., King, I., & Lyu, M. R. (2008, 15-19 Dec. 2008). *Formal Models for Expert Finding on DBLP Bibliography Data.* Paper presented at the 2008 Eighth IEEE International Conference on Data Mining.

Flick, U. (2008). *Designing Qualitative Research*: SAGE Publications.

Girvan, C., & Savage, T. (2013). Guidelines for Conducting Text Based Interviews in Virtual Worlds. In M. Childs & A. Peachey (Eds.), *Understanding Learning in Virtual Worlds* (pp. 21-39). London: Springer London.

Gretsch, S., Mandl, H., & Hense, J. (2011). *The Difficulty of Finding Experts-Implementation Process of Corporate Yellow Pages*. Paper presented at the KMIS.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569-16572. doi:10.1073/pnas.0507655102

Imhoff, C., & White, C. (2011). Self-service business intelligence: Empowering users to generate insights. *TDWI best practices report*, 40.

Isaak, D. (2016). PubMed2XL (version 2.01). *Journal of the Medical Library Association: JMLA, 104*(1), 92.

Kao, W.-C., Liu, D.-R., & Wang, S.-W. (2010). *Expert finding in question-answering websites: a novel hybrid approach*. Paper presented at the Proceedings of the 2010 ACM Symposium on Applied Computing, Sierre, Switzerland.

Kardan, A., Omidvar, A., & Farahmandnia, F. (2011, 17-19 May 2011). *Expert finding on social network with link analysis approach*. Paper presented at the 2011 19th Iranian Conference on Electrical Engineering.

Lin, S., Hong, W., Wang, D., & Li, T. (2017). A survey on expert finding techniques. *Journal of Intelligent Information Systems, 49*(2), 255-279. doi:10.1007/s10844-016-0440-5

Louise Barriball, K., & While, A. (1994). Collecting data using a semi-structured interview: a discussion paper. *Journal of Advanced Nursing, 19*(2), 328-335. doi:doi:10.1111/j.1365-2648.1994.tb01088.x

Markham, A. N. (2004). Internet communication as a tool for qualitative research. *Qualitative research: Theory, method and practice, 2*, 95-124.

Marshall, M. N. (1996). Sampling for qualitative research. *Family practice, 13*(6), 522-526.

Matsuo, Y., & Ishizuka, M. (2004). Keyword Extraction From A Single Document Using Word Co-Occurrence Statistical Information. *International Journal on Artificial Intelligence Tools, 13*(01), 157-169. doi:10.1142/s0218213004001466

Merrouni, Z. A., Frikh, B., & Ouhbi, B. (2016, 24-26 Oct. 2016). *Automatic keyphrase extraction: An overview of the state of the art.* Paper presented at the 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt).

Naeem, M., Khan, M. B., & Afzal, M. T. (2013). Expert Discovery: A web mining approach. *Journal of AI and Data Mining, 1*(1), 35-47. doi:10.22044/jadm.2013.116

Negash, S. (2004). Business intelligence. *Communications of the association for information systems, 13*(1), 177-195.

OpenRefine. (2018). Retrieved from http://openrefine.org/

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful Sampling for Qualitative

Data Collection and Analysis in Mixed Method Implementation Research. *Administration and Policy in Mental Health and Mental Health Services Research, 42*(5), 533-544. doi:10.1007/s10488-013-0528-y

Patton, M. Q. (2002). *Qualitative research and evaluation methods*. Thousand Oaks, CA: Sage.

Quatrini Carvalho Passos Guimarães, H. C., Pena, S. B., Lopes, J. d. L., Lopes, C. T., & Bottura Leite de Barros, A. L. (2016). Experts for Validation Studies in Nursing: New Proposal and Selection Criteria. *International Journal of Nursing Knowledge, 27*(3), 130-135. doi:doi:10.1111/2047-3095.12089

Riahi, F., Zolaktaf, Z., Shafiei, M., & Milios, E. (2012). *Finding expert users in community question answering*. Paper presented at the Proceedings of the 21st International Conference on World Wide Web, Lyon, France.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, 1-20.

Ru, Z., Xu, W., & Guo, J. (2007). *An Expert Experience Probabilistic Model for Enterprise Expert Finding*. Paper presented at the Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007).

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology, 15*(3), 234-281.

Saaty, T. L. (1990). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research, 48*(1), 9-26.

Saaty, T. L. (2013). Analytic Hierarchy Process. In S. I. Gass & M. C. Fu (Eds.), *Encyclopedia of Operations Research and Management Science* (pp. 52-64). Boston, MA: Springer US.

Serdyukov, P., & Hiemstra, D. (2008). *Modeling documents as mixtures of persons for expert finding*. Paper presented at the European Conference on Information Retrieval.

Sherman, R. (2014). *Business Intelligence Guidebook: From Data Integration to Analytics*. Newnes.

Smuts, M., Scholtz, B., & Calitz, A. (2015). *Design guidelines for business intelligence tools for novice users*. Paper presented at the Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists and Information Technologists.

SuperDecsion. (2018). Retrieved from http://www.superdecisions.com/

Verborgh, R., & De Wilde, M. (2013). *Using OpenRefine*: Packt Publishing Ltd.

Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W., & Zhang, Z. (2013). ExpertRank: A topic-aware expert finding algorithm for online knowledge communities. *Decision support systems, 54*(3), 1442-1451.

Yang, K.-H., Chen, C.-Y., Lee, H.-M., & Ho, J.-M. (2008). *EFS: Expert finding system based on Wikipedia link pattern analysis*. Paper presented at the 2008 IEEE International Conference on Systems, Man and Cybernetics.

Zhang, J., Tang, J., & Li, J. (2007). *Expert Finding in a Social Network*, Berlin, Heidelberg.

Zhang, J., Tang, J., Liu, L., & Li, J. (2008). *A Mixture Model for Expert Finding*, Berlin, Heidelberg.

Zhao, Z., Zhang, L., He, X., & Ng, W. (2014). Expert finding for question answering via graph regularized matrix completion. *IEEE Transactions on Knowledge and Data Engineering, 27*(4), 993-1004.