**POLYTECHNIQUE MONTRÉAL**

affiliée à l'Université de Montréal

# Exploring Innovation Concepts in Twitter Via LDA Topic Modelling the Case Custom Computer Programming Services

**MELIKA JAFARI**

Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie industriel

Décembre 2020

© Melika Jafari, 2020.

**POLYTECHNIQUE MONTRÉAL**

affiliée à l'Université de Montréal

Ce mémoire intitulé :

# Exploring Innovation Concepts in Twitter Via LDA Topic Modelling the Case Custom Computer Programming Services

présenté par **Melika Jafari**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

**Fabiano ARMELLINI,** président

**Catherine BEAUDRY,** membre et directrice de recherche

**Nathalie DE MARCELLIS-WARIN**, membre

# DEDICATION

***TO Mahdi-Alsaheb Alzaman-A.S-***

***The messenger of peace and tranquility.***

# ACKNOWLEDGEMENTS

I thank my supervisor M$^{me}$ Catherine Beaudry for accepting me as a master's student and for guiding me during these two years. I would like to thank my family for their support and friends and colleagues Davide Pulizzotto, Anas Ramdani, Faeze Vahdati ,Alvar Herrera, and Cintia.Blanco.

I would like to thank Mr. Fabiano Armellini and M$^{me}$. Nathalie Marcellis-Warinfor having devoted their time and expertise in composing the evaluation panel for this thesis.

I would like to thank my friends from Poly and the students of the Chair for all these moments lived in Montreal.

# RÉSUMÉ

Twitter, en tant qu'outil de micro-blogging, est l'une des plus célèbres plateformes de réseaux sociaux. Avec 500 millions de tweets par jour, c'est une source de données propice à l'exploration des messages publiés pour analyser le comportement et la communication en ligne des entreprises en termes de concepts d'innovation. Les différents types de contextes non structurés dans les postes présentent de nombreux défis. Des chercheurs suggèrent que de nouvelles approches d'apprentissage machine peuvent les aider à analyser ces données. Malgré les différents efforts des chercheurs pour analyser ce type de données, la littérature scientifique et opérationnelle n'a pas encore fourni de cadre d'analyse des concepts d'innovation par l'exploration de texte et le traitement du langage naturel (TLN) sur Twitter.

Ce travail de recherche vise à explorer si les entreprises des secteurs des services de programmation informatique personnalisée mentionnent différents concepts d'innovation dans leurs comptes Twitter, et si une utilisation simple et plutôt grossière du TNL peut rapidement prendre cela en compte. Plus précisément, l'objectif de cette étude est de répondre à trois questions de recherche : (1) dans quelle mesure les entreprises s'intéressent-elles aux concepts d'innovation dans leurs tweets ? (2) quels sont les mots les plus fréquemment utilisés dans les tweets des entreprises de ce secteur ? et enfin, (3) dans quelle mesure l'utilisation de l'algorithme « Latent Dirichlet Allocation » (LDA) identifie-t-elle efficacement les sujets liés à l'innovation dans les tweets des entreprises ?

Pour atteindre ces objectifs, du text mining et du modèle Latent Dirichlet Allocation (LDA) a été utilisée pour rechercher et explorer le contenu des tweets en ce qui concerne l'évaluation et la visualisation des mots fréquents et des sujets discutés dans les tweets. Nous avons spécifiquement fouillé le texte contenu dans les tweets pour trouver les mots liés à des concepts d'innovation spécifiques.

Dans un premier temps, cinq facteurs d'innovation ont été choisis et les mots clés associés provenant de différentes sources ont été rassemblés dans cinq tableaux de référence. Ces facteurs sont les suivants R&D, propriété intellectuelle, collaboration, financement externe et créativité, Pour la deuxième étape, l'utilisation de techniques de fouille de texte via LINQ et C# en studio visuel aide à répondre à la première question de recherche ; les résultats montrent que la

collaboration est le facteur d’innovation le plus utilisé dans les tweets des entreprises. En ce qui concerne la réponse aux deuxième et troisième questions de recherche, l’utilisation des techniques du TNL et la modélisation thématique de LDA permettent de trouver les termes les plus fréquents. De plus, en utilisant cinq facteurs d'innovation et en élargissant nos mots-clés liés à chaque facteur, les résultats montrent que le modèle LDA peut aider à trouver les sujets les plus probables dans les tweets, tels que la créativité, la R&D et la collaboration.

# ABSTRACT

As a micro-blogging tool, Twitter is one of the most popular social networking platforms. With 500 million tweets per day, it is an appropriate data source for mining broadcast messages to analyze firms' online behaviour and communication in terms of innovation concepts. Different types of unstructured contexts in the posts present many challenges, but researchers believe that new machine learning approaches can help them analyze these data. Despite the efforts of numerous researchers to investigate these kinds of data, the scientific and operational literature have not yet provided a proper framework for analyzing innovation concepts via text mining and natural language processing (NLP) on Twitter.

This research aims to explore whether firms in the Custom Computer Programming Services sector mention innovation concepts in their Twitter accounts and whether a simple and rather crude use of NLP can quickly pick this up. More specifically, this study aims to answer three research questions: (1) To what extent do firms demonstrate an interest in innovation concepts in their tweets? (2) Which words are used most frequently by firms in that sector? Finally, (3) to what extent does the use of Latent Dirichlet Allocation (LDA) effectively identify tweets related to innovation?

Text mining and the LDA model have been used to search and explore the tweet content to evaluate and visualize frequently used words. We mined tweet text to find words related to specific innovation concepts.

Our first step was to choose five innovation factors. They are R&D, IP (Intellectual Property), collaboration, external financing, and creativity. Also, we selected related keywords identified from different sources for these factors in five reference tables. For the second step, we used text mining techniques via LINQ and C# in Visual Studio to answer the first research question.

The results show that collaboration concepts are the most used in firms’ tweets. To answer the second and third research questions, we employed NLP techniques and LDA topic modelling using Python programming language to find the most frequent terms. Moreover, by using five innovation factors and expanding their related keywords results show that the LDA model can help find the most probable tweets as creativity, R&D and collaboration.

# TABLE OF CONTENTS

DEDICATION ..... III

ACKNOWLEDGEMENTS ..... IV

RÉSUMÉ ..... V

ABSTRACT ..... VII

TABLE OF CONTENTS ..... VIII

LIST OF TABLES ..... X

LIST OF FIGURES ..... XII

LIST OF SYMBOLS AND ABBREVIATIONS ..... XIV

LIST OF APPENDICES ..... XVI

CHAPTER 1 INTRODUCTION ..... 1

CHAPTER 2 LITERATURE REVIEW ..... 5

2.1 Text mining and analysis techniques applied to social media content ..... 5

2.1.1 Internet, Web2.0 and social media ..... 5

2.1.2 Big data analytics ..... 11

2.2 Structuring the innovation field for text mining purposes ..... 14

CHAPTER 3 METHODOLOGY ..... 33

3.1 Problem statement ..... 33

3.2 Research objectives and contributions ..... 34

3.3 Research framework ..... 34

3.4 Research data ..... 35

CHAPTER 4 RESULTS ..... 45

4.1 Innovation concepts and LDA topic modelling results ..... 45

4.2 Analyzing implementation of model and innovation factors as research findings ..... 57

CHAPTER 5 DISCUSSION ..... 62

5.1 Social media research challenges ..... 62

5.2 Discussion about the second and third research question ..... 64

CHAPTER 6 CONCLUSION AND RECOMMENDATIONS ..... 67

6.1 Research limitations ..... 67

6.2 Future studies ..... 69

REFERENCES ..... 70

APPENDICES ..... 83

# LIST OF TABLES

Table 2.1 Social media types and examples source (Liu et al., 2016)..... 6

Table 2.2 Classification of social media by social presence/media richness and self-presentation/self-disclosure source: (Kaplan & Haenlein, 2010)..... 6

Table 2.3 Factors and keywords about innovation concept, source: (Beaudry et al., 2016; Héroux-Vaillancourt et al., 2020)..... 17

Table 2.4 R&D and relevant keywords and themes (Factor 1)..... 20

Table 2.5 IP and relevant keywords and themes (Factor 2)..... 23

Table 2.6 Collaboration and relevant keywords and themes (Factor 3)..... 26

Table 2.7 External financing and relevant keywords and themes (Factor 4)..... 28

Table 2.8 Creativity and relevant keywords and themes (Factor 5)..... 31

Table 3.1 Data set variables..... 38

Table 3.2 LDA hyperparameters..... 43

Table 4.1 Number of times of using innovation factors and their keywords source: Héroux-Vaillancourt et al. (2020)..... 46

Table 4.2 Results of topic one by focusing on innovation factors..... 49

Table 4.3 Results of topic two by focusing on innovation factors..... 50

Table 4.4 Results of topic three by focusing on innovation factors..... 52

Table 4.5 Results of topic four by focusing on innovation factors..... 53

Table 4.6 Results of topic five by focusing on innovation factors..... 55

Table 4.7 Categorizing topics in dataset..... 57

Table 4.8 Comparing different number of topics based on their log-likelihood and perplexity... 61

Table 4.9 Model information..... 61

Table 5.1 Innovation words and themes in data set.....66

Table A.1 Information related to the code 514511 and two sub-divisions 511210 and 514512 ...83

Table A.2 Information about code 514512.....84

# LIST OF FIGURES

Figure 2.1 Why we use social networks? source :(Shao, 2009, p. 25) ..... 8

Figure 2.2 Three Vs of big data source: (Russom, 2011) ..... 11

Figure 3.1 Research framework ..... 34

Figure 3.2 Information of the percentage of having home page in the dataset for codes NAICS 514512 and 511210 ..... 36

Figure 3.3 Total homepage link in the dataset for codes NAICS 514512 and 511210 ..... 36

Figure 3.4 Percentages of using the Twitter account for codes NAICS 514512 and 511210 ..... 37

Figure 3.5 A LDA graphical model source : (David M Blei et al., 2003) ..... 42

Figure 4.1 The 27 most salient terms in data set ..... 47

Figure 4.2 Relevant terms for topic one ..... 49

Figure 4.3 Most relevant terms for topic two ..... 51

Figure 4.4 Most relevant terms for topic three ..... 52

Figure 4.5 Most relevant terms for topic four ..... 54

Figure 4.6 Most relevant terms for topic five ..... 56

Figure 4.7 Implementing LDA with different number of topics ..... 60

Figure 4.8 Log-likelihood and perplexity based on different number of topics ..... 60

Figure 4.9 The best model for dataset ..... 61

Figure B.1 Codes for collecting companies' information ..... 85

Figure B.2 A JSON file sample for a company ..... 85

Figure B.3 Importing libraries and data set in the Jupiter notebook ..... 86

Figure B.4 Data description ..... 86

Figure B.5 Data description ..... 87

Figure B.6 Information about the size of data set ..... 87

Figure B.7 Removing punctuation function and its result ..... 87

Figure B.8 Tokenization function and its result ..... 88

Figure B.9 Using NLTK and removing stop words function and its result ..... 88

Figure B.10 Stemming the text and Its result ..... 89

Figure B.11 Lemmatizing the text and its result ..... 89

Figure B.12 Cleaning the text and count vectorising it ..... 89

# LIST OF SYMBOLS AND ABBREVIATIONS

| | |
|---|---|
| AI: | Artificial intelligence |
| ATDC: | Aberdeen Technology Data Cloud |
| BDA: | Big data analytics |
| CIS: | Community Innovation Survey |
| CTM: | Correlated Topic Model |
| IoT: | Internet of things |
| IP: | Intellectual Property |
| IR: | Information retrieval |
| K-NN: | K-Nearest Neighbor |
| LAN: | Local Area Network |
| LDA: | Latent Dirichlet Allocation |
| LINQ: | Language Integrated Query |
| LLSF: | Linear Least Square Fit |
| LSA: | Latent Semantic Analysis |
| NAICS: | North American Industry Classification System |
| NB: | Native Bayes |
| NLP: | Natural language processing |
| NLTK: | Natural Language Toolkit |
| OECD: | Organization for Economic Co-operation and Development |
| PLSA: | Probabilistic Latent Semantic Analysis |
| R&D: | Research and development |
| RT: | Retweet |
| RTD: | Research Technological development |
| SMEs: | Small and medium-sized enterprises |
| SNA: | Social network analysis |
| SVM: | Support Vector Machines |
| TF-IDF: | Term Frequency-Inverse Document Frequency |
| VSM: | Vector Space Module |

| | |
|---|---|
| XML: | Extensible Markup Language |
| #: | Hashtag |
| @: | Mention |

# LIST OF APPENDICES

APPENDIX A NAICS CODES INFORMATION ..... 83

APPENDIX B TOPIC MODELLING AND TEXT MINING ..... 85

# CHAPTER 1 INTRODUCTION

Social networks platforms such as Twitter, Facebook, and Tumbler can create, generate, and save users' data and information, providing significant awareness regarding the firms' online behaviors. Twitter, as a micro-blogging platform, picks up colossal ubiquitous details. Its public accessibility offers an outstanding opportunity for researchers to mine and analyze its content.

This research project will apply text mining and natural language processing (NLP) to explore the content of Twitter data. The technical definition for the NLP, is "a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis to achieve human-like language processing for a range of tasks or applications" (Liddy, 2001a, p. 2). These analyses are performed through several basic steps such as sentence segmentation, word tokenization, and several other measures. Social networks texts like those found on Twitter represent a collection of different information. These texts are unstructured, written with different formats by multiple people in various languages, styles, and cultures with different purposes. Some research shows that using Twitter in industries can improve the firms' performances or that this social platform constitutes an influential network of relations between industrial partners (Guercini, Misopoulos, Mitic, Kapoulas, & Karapiperis, 2014; Leung, Bai, & Stahura, 2015; Punel & Ermagun, 2018). Some research domains exploit Twitter data that focuses on firms' text and concepts (X. Liu, Burns, & Hou, 2017; Luo, Zhang, & Duan, 2013; Shen, Luong, Ho, & Djailani, 2020). Other studies focus on online web pages to examine innovation concepts (Ramis Ferrer, & Luis Martinez Lastra, 2019; Lim & Maglio, 2018; Riasanow, Jäntgen, Hermes, Böhm, & Krcmar, 2020). For example, Gök, Waterworth, and Shapira (2015) focused on 296 green goods SMEs' web pages to mine online keywords related to R&D activities. Héroux-Vaillancourt et al. (2020) used content mining techniques focused on four innovation factors which are R&D, Intellectual property, collaboration, and external financing Canadian nanotechnology company webpages. To analyze innovation text, the study by Du Plessis (2007) on using text mining methods showed that innovation process could have six different attributes, proving that an innovation process uses a different definition for each attribute. Consequently, few research articles and published firms' cases on using innovation concepts and social media, such as Twitter, use NLP algorithms and LDA topic modelling. For example, due to understanding tweet topics, the topic summarization research domain uses LDA (O'Connor, Krieger, & Ahn, 2010). Additionally,

Mehrotra, Sanner, Buntine, and Xie (2013) showed that tweet pooling improves the measures for the topics found using LDA. He, Jia, Han, and Ding (2014) explained that using topic modeling like LDA can discover users interest in microblogs like Twitter.

**Importance of the study**

Every industrial revolution brings challenges and opportunities. For instance, in the first industrial revolution, the steam engine developed communications and transportation in Great Britain. As an example of the second industrial revolution, telephone communication in the Unites States was associated with several benefits (Carr, 2003). In the third industrial revolution, the Internet plays a key role in transforming the world economic landscape. These revolutions all had positive results regarding economic growth, increased productivities, etc. With the advent of Industry 4.0, which includes the financial, social, and environmental aspects of firms as a "smart factory", communication has tremendously increased via, for instance, the Internet and related services (Buhr, 2015; Dutton, 2014). Besides, according to Chaffey (2016), with the world's total population at around seven and a half billion people, statistics show that in the year 2020, there are more than four and a half billion Internet users (IOS, 2020).

As one of the electronic communication platforms, social networks expanded throughout the world for both individual and global communities. They now count nearly four billion active people and organizations (IOS, 2020). This platform allows users to generate and share various types of information or content, such as words, pictures, videos, and audio. The use of social networks by individuals and firms alike in this highly competitive era has contributed to change the firms' industrial and business processes. Collecting and analyzing the Internet's content in terms of knowledge and data regarding new concepts at the core of the so-called industry 4.0, such as cloud, social networks, social media, big data, and the Internet of Things (IoT), is therefore timely. It also provides opportunities for organizations, companies, and even governments to gain extensive information about the various dimensions of their lives (Geiger & Sá, 2008). Therefore, the analysis of social media has imposed itself as one important area for research based on new methods such as artificial intelligence(AI), natural language processing (NLP), social network analysis (SNA), and big data analytics (BDA) (Peters, Chen, Kaplan, Ognibeni, & Pauwels, 2013; Yaqoob et al., 2016).

Nowadays, establishing, using, managing, and supporting social media is both a professional and a work area. This new revolution needs computer system integration and information management knowledge to develop application and publish software skills. Innovation can be defined as applying new ideas for the products, processes, or other aspects of the activities of an organization that lead to increased value and meeting consumers' needs and desires (Cancino, Merigó, & Palacios-Marqués, 2015). Therefore, it is fundamental to developing a country's economy (Zhu & Guan, 2012). According to Mowery and Oxley (1995), technological advances in innovations play a leading role in helping countries achieve their competitive advantages.

Based on Niosi (2000), Canadian firms are strong in biotechnology, software development, and telecommunication equipment, and the country's research focus is primarily on small firms. Therefore, digitalization and the use of innovative computer technologies, which can be as innovation process, are increasingly popular and it can be applied to each individual, industrial, and business unites to improve their productivity. As technology is a critical resource of the firm and business growth in different industries, its use in the innovation concepts as an element that influences relationships among other actors is a widely discussed topic (Adner, 2006; Adner & Kapoor, 2010).

In this research, Custom Computer Programming Services sectors (NAICS 541511) as research data set has been selected. The professional domain of these firms is related to establishing, publishing, installing, and supporting their products and services. Innovation has a vital role in their professional work life. Focusing on their social media and especially their Twitter accounts, in terms of using innovation concepts itself can be a source of improvement in these business groups.

Besides, some uncertain situation like living in COVID -19 condition, changed firms' professional lifestyle in the aspect of using online platforms, software, websites, applications, local area network(LANs) and computer equipment forced firms to start using technology or to upgrade their existing systems.

This study's goal is related to the using innovation concepts on firms' Twitter accounts. We used a database of Information and Communication Technologies (ICT) adoption to identify Canadian firms' web sites and analysing 405 firms with active Twitter accounts.

For this research, we collected keywords from five innovation factors. Different resources, such as Héroux-Vaillancourt et al. (2020), and the older Oslo Manual OECD (2005), O. Manual (2018) and glossary of statistical terms web page by OECD, Williams et al. (2016) and Collins thesaurus of the English language (in order to know slangs and informal language used in tweets) were selected. These five factors were: R&D, IP, collaboration, external financing, and creativity. For each factor, there were different keywords and themes which saved in individual tables and called reference tables. Therefore, having five reference tables of these words and using text mining techniques via LINQ and C# in Visual Studio helped us evaluate firms' interests regarding innovation factors. Furthermore, focusing on these reference tables and using LDA topic modelling enabled us to find most frequently used and guided us to predict which of the five innovation factors would be a probable topic for tweets.

The rest of thesis is organized as follows: Chapter two reviews the prior literature and it covers the research related to social media networks analysis, with a primary on the Twitter platforms and text mining. These subjects build the research's conceptual framework. Chapter three presents the study's research questions and objectives and the methodology used related to the NLP's topic modeling. The fourth chapter presents text mining and LDA topic modelling results in Twitter data based on five innovation factors. Chapter five discusses the general findings of research questions and some challenges; the thesis concludes with chapter six, that unfold the thesis recommendation for research limitation and research future work.

# CHAPTER 2 LITERATURE REVIEW

This research involves using LDA topic modelling concerning innovation concepts. Before doing so, this chapter starts with presenting in section 2.1 a literature review on the work done related to the Internet, Web2.0, and social media. Section 2.2 looks at innovation concepts that are present in the collected data from Twitter accounts of the firms.

## 2.1 Text mining and analysis techniques applied to social media content

### 2.1.1 Internet, Web2.0 and social media

The invention of the World Wide Web by Tim Berners-Lee in March 1989 and the third industrial revolution have changed people's lifestyles in terms of online communication aspects. According to Beal (2010)'s idea , the World Wide Web is an easy way of accessing information-sharing via internet. According to Constantinides and Fountain (2008), using Web 2.0 as an open-source platform can help businesses and social participants share or edit their information , knowledge ,and experiences. Therefore, Web 2.0 platforms are used by every individual social media user (Kaplan & Haenlein, 2010). According to Constantinides and Fountain (2008), there are some categories for the web 2.0 such as 'blogs,' 'social networks,' 'content communities,' 'forums,' and 'content aggregators.' O'Reilly and Battelle (2009) mention that social media is one of the most valuable aspects of Web 2.0 which can be widespread and has Web 2.0 features. Regarding D. Evans (2010), Web 2.0 conversations, social interactions, group formations, and social media features are related to the realization of concepts and context apprehension in the conversation. According to Xiang and Gretzel (2010), there are some gaps in the literature and understanding of the definition of social media and some scholars believe that social media represents behavioural and online activities of users with the aim of sharing information and knowledge in communication aspects (Kaplan & Haenlein, 2010; Safko, 2010; Xiang & Gretzel, 2010). Some scholars agree that two other features of social media are time and interactive attributes (Bortree & Seltzer, 2009; Perry, Taylor, & Doerfel, 2003; Wright & Hinson, 2009).

**Classification of social media**

Based on Kietzmann, Hermkens, McCarthy, and Silvestre (2011), seven functional blocks for social media can be defined as follows : (1) 'identity,' (2) 'conversations,' (3) 'sharing,' (4)

‘presence,’ (5) ‘relationships,’ (6) ‘reputation,’ and (7) ‘groups.’ There are different types of social media based on their reason of activities. Table 2.1 illustrates some social media types and examples regard to Liu, Fraustino, and Jin (2016).

Table 2.1 Social media types and examples source (Liu et al., 2016)

| **Social media type** | **Examples** |
|---|---|
| Blogs | Blogger, WordPress |
| Discussion Forums | LiveJournal, ProBoards |
| Micro-blogs | Tumblr, Twitter |
| Photo/Video Sharing & Podcasting | Flickr, iTunes Podcasts, YouTube, Pinterest |
| Social Bookmarking | Del.icio.us, Diigo |
| Social Discovery Engines & News Sources | Reddit, StumbleUpon, Slashdot |
| Social/Professional Networking | Facebook, Google+, LinkedIn, MySpace |
| Social Rating/Reviews | AngiesList, Yelp |
| Video/Text Chatting | Skype, AIM, mobile texting |
| Wikis | Wikipedia, Wikispaces |

According to Kaplan and Haenlein (2010), classification of social media can be based on presence/ media richness and self -presentation/ self-disclosure. Table 2.2 presents these classifications

Table 2.2 Classification of social media by social presence/media richness and self-presentation/self-disclosure source: (Kaplan & Haenlein, 2010)

| | | **Social presence/Media richness** | | |
|---|---|---|---|---|
| | | Low | Medium | High |
| Self-presentation / self-disclosure | High | Blogs | Social network sites (e.g., Facebook) | Virtual social worlds (e.g., Second life, IMVU) |
| | Low | Collaborative projects (e.g., Wikipedia) | Content communities (e.g., YouTube) | Virtual game worlds (e.g., Asgard’s Wrath and World of Warcraft) |

Since the concepts of social media and social networks are sometime interchangeable, scholars such as Kietzmann et al. (2011) and Leonardi, Huysman, and Steinfield (2013) consider social

media as a channel of communication in an organization. It has two parties: external parties and internal parties which allows users (both individual and organizations) to communicate, share and exchange data and information and this information can be divided in different formats such as text, video and images, while focusing on the users' social connections and interaction can be referred to the social networks. Hence, by prioritizing social connection and networking, our research defines social networks as an online application of social media and other social media functions, including their services. Therefore, unavoidably, social media is not a social network because some of social media's primary goals are focused on sharing information and not on networking (e.g., YouTube).

**Application of social media in firms**

There are three reasons users create content and make it available for other users on social network sites: (1) to answer the needs of information, entertainment and their mood; (2) to connect with others; and (3) to show their self-actualization and self-expression (Vickery & Wunsch-Vincent, 2007). "Interdependence of people's consuming, participating, and producing on user-generated media." Will be shown in Figure 2.1 by Shao (2009, p. 25).

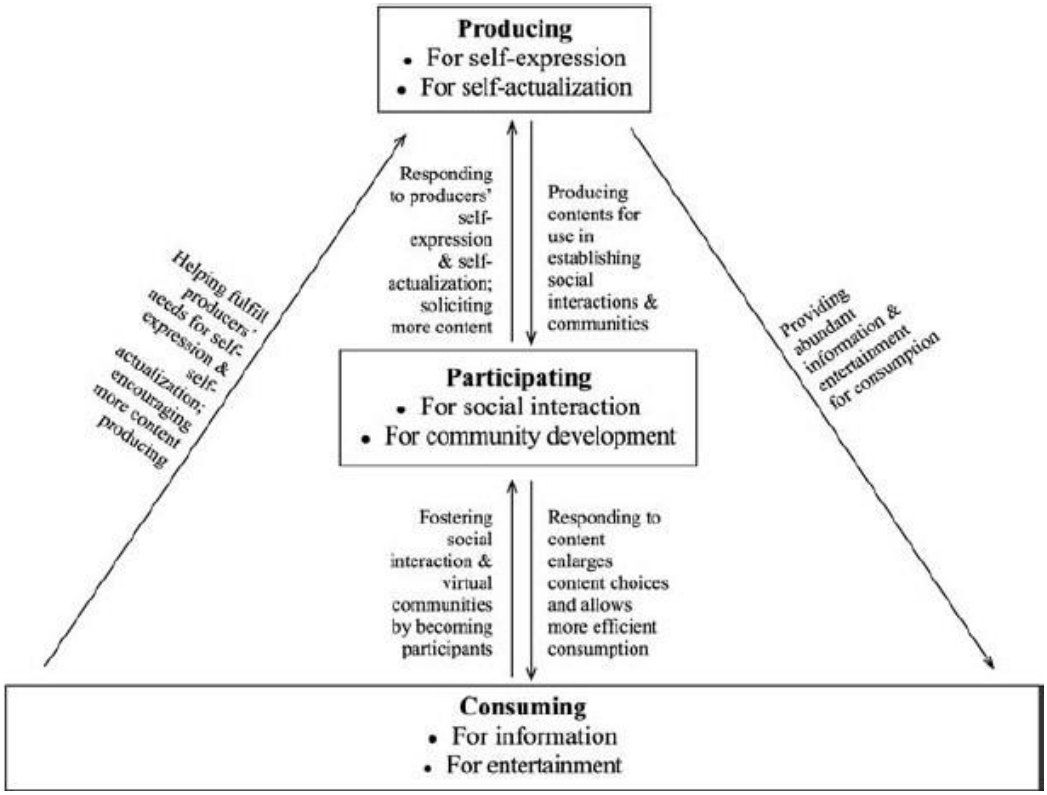


Figure 2.1 Why we use social networks? source :(Shao, 2009, p. 25)

According to Whiting and Williams (2013, p. 364), users utilize social networks for seven reasons : "(1)social interaction, (2) information seeking, (3) pass time, (4) entertainment, (5) relaxation,(6) communicator utility and (7) convenience utility."

With respect to research from Potts, Cunningham, Hartley, and Ormerod (2008) , using social networks in different industries can offer them some opportunities such as creativity in their market. According to Culnan, McHugh, and Zubillaga (2010), social media platforms in industries as a means of communication and collecting users' information that provide an improvement opportunity for all firms. Concerning the different kinds of social media (e.g., Facebook, Twitter) used by organizations of various sizes many researchers show how much these online platforms can help industries for their future (W. M. Campbell, Dagli, & Weinstein, 2013; Lieberman, 2014; Michalak, Rahwan, & Wooldridge, 2017). Focusing on subjects such as collaboration, innovation, and the use of social media platforms in further research showed that there is a significant and positive relationship between product innovation and feedback by users on social media. Besides,

the Internet plays the primary role in the process of collaboration innovation. Social media can facilitate collaborative innovation in the product development process and provide a place for collaboration between firms and users, especially influencers (Bertschek & Kesler, 2017; Bolstad & Høili, 2019; Sawhney, Verona, & Prandelli, 2005). Focusing on the market and branding , De Veirman, Cauberghe, and Hudders (2017) discussed that the use of social media (in this case, Instagram) resulted in users and influencers helping firms' market branding.

Contents that are created by users in online platforms can affect on firms' performance. For example, according to research by Goh, Heng, and Lin (2013) , data from Facebook users ( consumers of the firms' product) affects consumer purchase behaviours. Aspasia and Ourania (2015) discussed that analyzing customers' behaviours based on Facebook post can provide rich information useful for responding consumers' messages. Based on Chung, Animesh, Han, and Pinsonneault (2014), social media can help firms in their business value and performance .Tajvidi and Karami (2017), by focusing on marketing capabilities' role , found that branding and innovation have positive and significant between social media use and firm performance. According to Culnan et al. (2010), social media could help large firms incorporate community and business value , they suggested firms' social media implementation guidelines. Concerning F. Wang and Vaughan (2014), the use of online platforms has appositive and significant relationship between a firm's marketing ( advertising efficiency ) and its web visibility. Moreover, there is a positive and significant relationship between firms' share value and web visibility. Focusing on Twitter activity and how it contributes firms' value, Majumdar and Bose (2019) discussed that Twitter has a substantial impact on firms' content and consequently, managers were encouraged to share information related to their products on Twitter and social media. Believing that social media has a significant impact on internal operations, marketing, and customer services, they also developed a model by focusing on some of the challenges of implementing social media for small businesses.

**Twitter and analyzing Twitter as a big data source**

Online social platforms provide an environment that allows users to show their behaviours easily. According to Majumdar and Bose (2019), users' behaviours interact with others through various interfaces. Users may upload and view content, choose friendships, and rank content as favourable, along with many other interactions.

Since 2005, the use of internet social networking platforms such as Twitter has increased. Twitter permits users to use "Twitter.com" or the application on their digital devices to submit their words as "tweets." The brevity of users' interactions imposed by the 140-character limit makes Twitter a unique place. Some symbols and options facilitate communication behaviours on Twitter, such as hashtags (#) and mentions (@).

According to Clark (2009), a hashtag (#) before a word or phrase helps users to link all of the other tweets that include that word or phrase. Moreover, it is useful for seeing the longevity of a conversation, recognizing tweets' categories, and finding concepts quickly. Hashtags can open the door to new communication pathways and increase awareness. Based on Kywe, Hoang, Lim, and Zhu (2012), personalized hashtags can be used as a simple method of collaborative filtering to improve the performance of tweet content. Regarding Cui, Zhang, Liu, Ma, and Zhang (2012, p. 1794), the hashtag plays an indicator role for an event. It has three attributes "(1) instability for temporal analysis, (2) Twitter meme possibility to distinguish social events from virtual topics or memes, and (3) authorship entropy for mining the most contributed authors."

According to Wikström (2014), hashtags can be a creative, communicative function for detecting information.

Weller, Bruns, Burgess, Mahrt, and Puschmann (2014), explain that a "mention", represented by the "@" sign, is another helpful symbol used on Twitter. Using this symbol helps to draw attention to another Twitter account. Moreover, it tries to represent significant challenges for the right audiences' correct information (A.-H. Tan, 1999b).

Several industrial studies showed that using the "@" symbol can be a good strategy for targeting audience and branding (Carbonell, Mayer, & Bravo, 2015; Shuai, Pepe, & Bollen, 2012; Yin, Fabbri, Rosenbloom, & Malin, 2015).

### 2.1.2 Big data analytics

In the 1990s, John Mashey first popularized the concept and use of the term “big data” (Lohr, 2013; Smaiti & Hanoune, 2015). According to Onay and Öztürk (2018, p. 382) the characteristics of big data are (1) high-Volume (related to the data size, (2) high-variety (type and nature of data) and (3) high- velocity (data generation speed). Figure 2.2 illustrates three Vs of big data.

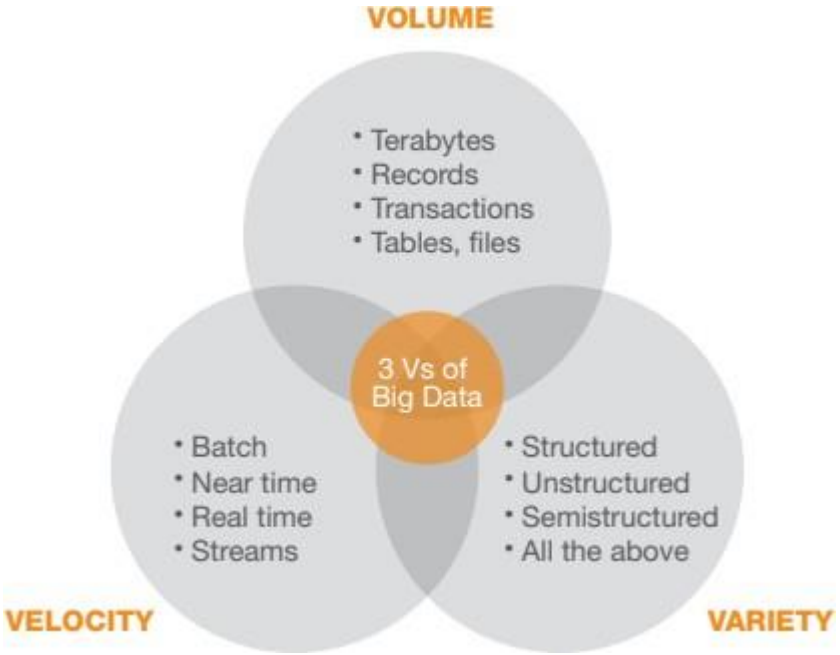


Figure 2.2 Three Vs of big data source: (Russom, 2011)

Regarding the Russom (2011, p. 7), the types of big data are structured data, unstructured data (text and human language), and semi-structured data (XML, RSS feeds). Other categories of data can be found in audio and video formats.

According to K. H. Tan, Zhan, Ji, Ye, and Chang (2015) , it is possible to extract new ideas or understanding about a firm’s products, customers, and market by accessing big data from social networks such as Twitter. Therefore, data can play an essential role in creating value and competitive advantages in firms. Based on the point of view presented by Driscoll and Walker (2014) , Twitter can be a data source that integrates into a visual presentation of events. Comparing Twitter with other social media platforms such as Facebook and YouTube shows that Twitter is

public enough. It does not need a high-bandwidth connection and is easily accessible (Driscoll & Walker, 2014).

Data mining and machine learning help scientists extract clear and unclear relationships among the billions of records of different types of social media data. Data mining research differs from statistical analysis. In statistical analysis, it is essential to have a hypothesis and to use empirical data to analyze the hypothesized relationship. In contrast, data mining does not need a hypothesis, and it can be used for different types of non-numerical data. According to Plotnikova (2018, p. 50), critical features for data mining are as follows: (1) it can automatically detect patterns, (2) it can predict the probability of results and outputs, (3) it can contain executive and helpful information, and (4) it can include extensive data and databases.

**Machine learning and big data**

Machine learning (ML) is one of the most outstanding human solutions to complex repetitive tasks. It is a sub-discipline of artificial intelligence (AI) and data mining. According to Bishop (2006), ML appears when a computer learns a specific function with unique algorithms and there is an application for it, such as image and voice detection, in user behaviour analytics. According to different studies, ML projects have eight crucial stages: (1) Data acquisition, (2) Data cleaning, (3) Feature extraction, (4) Train datasets, (5) Train ML models, (6) Test dataset, (7) Evaluate model and (8) Deploy model (Khan, Baharudin, Lee, & Khan, 2010; Mathiak & Eckstein, 2004). According to Sag (2019), the first and second steps of an ML project are related to data recognition and data preparation. The second step will repeat until the best results are found. ML can be divided into four different methods: (1) supervised learning (e.g., classification and regression); (2) unsupervised learning (e.g., clustering); (3) semi-supervised learning (e.g., a mixture of unsupervised and supervised methods), and (4) reinforcement learning (e.g., based on right and wrong results) (Bishop, 2006; Ratner & Ré, 2018; Tucker, 2004; Wiering & van Otterlo, 2012).

**Text mining, natural language processing and big data**

Text mining (text data mining or text analytics) is the qualitative analysis process that helps drive text information. Each text source, such as books, articles, websites, and social media knowledge, needs a specific approach to analyzing text. Therefore, there are three different perspectives of text mining: (1) information extraction, (2) data mining, and (3) knowledge discovery in the database

process (Hotho, Nürnberger, & Paaß, 2005). Text mining as a new technology of data mining (Berry & Kogan, 2010; Feldman & Sanger, 2007; Francis, 2006), was first proposed by Ronen Feldman (Berry & Kogan, 2010) to discover knowledge by extracting patterns in large collections of texts. The main text mining techniques include classification, clustering, summarization, information extraction, distribution analysis, and trend prediction. Classification methods in the text can be (1) Native Bayes (NB), (2) K-Nearest Neighbor (K-NN), (3) Support Vector Machines (SVM), (4) Vector Space Module (VSM), and (5) Linear Least Square Fit (LLSF) (Berry; Berry & Kogan, 2010; Francis, 2006; Franke, Nakhaeizadeh, & Renz, 2003; Karanikas & Theodoulidis, 2002; Song, 2008; Srivastava & Sahami, 2009; Zanasi, 2007).

According to Liddy (2001b), the computerized analysis approach called natural language processing (NLP) is based mainly on theories and a set of technologies. These texts can be both oral and written human communications. Because there are multiple types of language processing[1] and confusion in the text, numerous NLP systems can perform various levels of linguistic analysis. These systems can be in four categories: (1) paraphrasing an input text, (2) translating text between two different languages, (3) answering questions based on text contents, and (4) drawing inferences based on the text. The most common processing applications that use NLP to provide both theory and implementations are (1) information retrieval, (2) information extraction, (3) question answering, (4) summarization, (5) machine translation, and (6) dialogue systems.

Focusing on the industry's development based on computer sciences and analyzing concepts, researchers show that practical and valuable methods can help a firm overcome its challenges (Cuzzocrea, Loia, & Tommasetti, 2017; Iammarino, 2005; Jin, Wang, Chu, & Xia, 2018). There are different aggregation techniques for analyzing text on social media like Twitter. According to Java, Song, Finin, and Tseng (2007), firms save 80% of their information in text formats, such as research papers, news articles, webpages, e-mail, reports and social media. ML scholars in the field

---

[1] "Natural language processing approaches fall roughly into four categories: symbolic, statistical, connectionist, and hybrid" (Liddy, 2001a)

of text mining used some probabilistic topic modelling and unsupervised ML algorithms such as Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Indexing (PLSI), and Latent Dirichlet Allocation (LDA) to find hidden information automatically (David M Blei & Jordan, 2006; David M Blei, Ng, & Jordan, 2003b; Brants, 2005; Buntine, 2009; T Hofmann, 1999; Thomas Hofmann, 1999; Lu, Mei, & Zhai, 2011; Minka, 2013). One of the widely used models for text classification, text annotations, and text probabilistic topic modelling used in this study is the LDA model(David Meir Blei, 2004). Using LDA helps to understand probability distribution of topics and simplifies text generative process with large-scale text sets.

On Twitter text, obtaining topics with messages and their writer was discussed in some research (Ramage, Dumais, & Liebling, 2010; Weng, Lim, Jiang, & He, 2010). According to Hong and Davison (2010), training topic modelling with the LDA in Twitter could obtain a higher quality for better word classification performance. In other research into understanding Twitter messages, Hong and Davison (2010) tried to focus on the training model to compare the quality and effectiveness of classification problems. Weng et al. (2010) discussed about influential users on Twitter and the topical similarity between users and account structure. Asghari, Sierra-Sosa, and Elmaghraby (2018) in their study, used topic modelling on the tweets of healthcare to attempt to classify and label new tweets and improve the accuracy of predictions. According to Lansley and Longley (2016), using topic modelling algorithms and LDA, helps classify the geotagged tweets and shows that the nature of the content posted is based on the characteristics of places and users who send the post. In other research, Samarawickrama, Karunasekera, and Harwood (2015) focused on the evaluating different grouping schemes to find k high -level topic and finding words. The results showed that it is essential to group tweets together to achieve the best result.

## 2.2 Structuring the innovation field for text mining purposes

According to Chakraborty and Pagolu (2014), text mining by using computational methods and techniques can extract high-quality information from large amount of unstructured text. Moreover, it can seek to find especial concepts in the massive amount of information (Irfan, King et al. 2015). While a substantial amount of research is devoted to analyzing innovation concepts on firms' activities (e.g., products, market), the subject of using innovation concepts on firms' Twitter accounts - as a social media platform - via new techniques such as NLP remains largely under-

researched. Since this research focuses on innovation concepts and their keywords by implementing LDA topic modelling in tweets, it is essential to have a clear perspective on achieving better results.

**Innovation concepts keywords in online text mining**

Innovation can be defined as applying new ideas for products, processes, or other aspects of an organization's activities that lead to increased value and the ability to better meet consumer needs and desires (Cancino, Merigó et al. 2015). It is, therefore, a fundamental issue for the development of a country's economy (Lim and Maglio 2018, Cai, Ramis Ferrer et al. 2019, Riasanow, Jäntgen et al. 2020).

Studies related to national innovation systems illustrate a practical approach to innovation research that has influenced the innovative performance of firms and economies (Patel and Pavitt 1994, Lundvall 2016). Text mining methods in social media can provide a clear understanding of how to easily seek and find special concepts in the massive amount of information (Irfan, King et al. 2015). For instance, Salloum, Al-Emran et al. (2017) compared further research related to text mining techniques on Facebook and Twitter, as two critical social networks. According to Du Plessis (2007), text mining methods showed that the innovation process could have six different attributes, with a different definition for each point. These attributes are (1) stages (e.g., creation, generation, implementation, development, adoption); (2) social (e.g., organizations, firms, customers, social systems, employees, developers); (3) means (e.g., technology, ideas, inventions, creativity, market); (4) nature (e.g., new, improve, change); (5) type (e.g., product, service, process, technical); and (6) aim (e.g., succeed, differentiate, compete). Some of this research is focused on mining innovation of webpages. For example, another study focused on the highly innovative SMEs' websites to examine commercialisation-business model-related keywords (Libaers, Hicks, & Porter, 2016). The web content and web structure of nanotechnology websites were studied by Hyun Kim (2012) with regards to the "Triple Helix" (Etzkowitz & Leydesdorff, 2000). Moreover, web crawling techniques allows authors to search keywords across all webpages to draw conclusions on the SMEs' degree of innovation. For example, Gök, Waterworth, and Shapira (2015) focused on 296 SMEs in the green goods sector to find that online keyword mining provides additional insights compared to other more traditional ways of measuring R&D and invention through publications patents. In terms of innovation, the combination of obtrusive methods (e.g.,

firm innovation survey) and unobtrusive methods (e.g., patent database analyzing) can help researchers understand firms' activities better. For example, Beaudry, Héroux-Vaillancourt, and Rietsch (2016) focused on the websites of 89 nanotechnology firms and 71 aerospace firms in their research and, using content mining techniques to perform a word frequency analysis, showed that four essential factors (R&D, intellectual property, collaboration, and external financing) play prominent roles in the commercialization of high technology firms' webpages. Additionally, the use of web scrapers such as Nutch helped to extract and store firms' texts. Specifically, focusing on each core factor, some relevant keywords were collected from firms' webpages.

The same authors focused on the corporate websites of 79 Canadian nanotechnology and advanced material firms in subsequent research [2].These factors and their keywords were updated, and some additional keywords were added. Table 2.3 presents the keywords related to these four factors based on the two studies.

[2] Héroux-Vaillancourt, Beaudry, and Rietsch (2020)

Table 2.3 Factors and keywords about innovation concept, source: (Beaudry et al., 2016; Héroux-Vaillancourt et al., 2020)

| **Factors** | **Keywords** |
|---|---|
| R&D | research & development, research and development, r&d, researcher, product development, technology, development, technical development, development, phase, development program, development process, development project, development facility, technological development, development effort, development cycle, development research, development activity, fundamental research, basic research |
| Intellectual property (IP) | Patent, intellectual property, trade secret, industrial design |
| Collaboration | Affiliation, collaboration, cooperation, partners, Partnership, consort, international consortia, global consort |
| External financing | Atlantic Canada opportunities agency, business development bank of Canada, sustainable development technology, venture capital, Atlantic innovation fund, nrc-irap, fednor, Industrial research assistance program, grants, private investment |

In order to understand these factors and their keywords and use them in the current study, the next section explains how they are defined by different resources.

**Innovation concepts definitions**

Business sectors can measure their innovation based on the OECD/Eurostat (2018). In the older 2005 version of the manual (OECD, 2005), an experimental survey in Canada, Germany, Nordic countries, and the United States helped scholars learn about firms and their innovation on products and processes. This survey manual followed its 1997 version, which concentrated on the non-agricultural economy and considered that innovation accounted for different scopes, including organizational change and development of exciting markets. The new edition of Oslo manual OECD/Eurostat (2018), compared to the 2005 previous edition, provides practical guidance regarding indicators and quantitative analyses on innovation data. This provided a conceptual framework as well as some generally applicable updated innovation definitions for all sectors.

Moreover, this reference presented some information regarding internal and external factors on business innovation and measuring innovation in developing countries.

The new definition of innovation based on the 2018 Oslo manual is unlikely to have trickled down to the firms yet. The use of keywords presented in Table 2.3 and their use in this study, including in resources such as the older Oslo Manual OECD (2005) , O. Manual (2018) and glossary of statistical terms web page by OECD [3] can help to uncover more hidden knowledge or relevant keywords useful for applying LDA topic modelling results.

**Factor 1: R&D**

Due to changing the nature of R&D, several efforts have been made to explain this concept (Howells, 2008; Miles, 2007). In the seventh edition of OECD, R&D was denied as:

"*…creative work undertaken on a systematic basis in order to increase the stock of knowledge - including knowledge of humankind, culture and society - and to devise new applications of available knowledge*" Source *:* (Frascati, 2015, p. 44)

Therefore, according to the Frascati (2015, p. 44), R&D covers three types of activity: basic (fundamental) research, applied research, and experimental development. Basic research is "theoretical or experimental work undertaken primarily to acquire new knowledge of the underlying foundations of phenomena and observable facts, without any particular application or use." Scientists do basic research, and these results are published in scientific journals. Applied research is "original investigation to acquire new knowledge and is directed primarily towards a specific practical objective." Applied research can be exploring basic research findings or applying new methods or ways specific to research objectives. Experimental development is "systematic work, drawing on existing knowledge obtained from research and experience, which is directed to producing new materials, products and devices; to installing new systems, processes, and services; or to improving those already installed or produced." (Frascati, 2015, p. 45). To generating new knowledge, there are some activities that have to be done by R&D. These activities can be categorized by R&D projects (Frascati, 2015).

---

[3] The OECD Glossary of Statistical Terms

According to Frascati (2015), five criteria are used to identify R&D activities: (1) novel, (2) creative, (3) uncertain, (4) systematic, and finally, (5) transferable. These five criteria help to distinguish R&D from the other parts of the innovation process.

In European countries, research and development (R&D, R+D) is called research and technological development (RTD)[4]. It is related to all innovative activities done by firms, governments, and corporations, which results in the development or improvement of existing or developing services and products (Kassema, 2019). Expenditures on R&D and the number of employees devoted to the activity can be a good measure of R&D efforts. There are some advantages and disadvantages. Since the 1950s, R&D data has been collected and compiled by the OECD. Dividing R&D effort based on the product (not the process) can help measure firms' performance, firms' growth, and the rate of employment levels and profits. Relevant themes, keywords and concepts related to R&D from different resources will illustrate in Table 2.4.

[4] https://en.wikipedia.org/wiki/Research_and_development

Table 2.4 R&D and relevant keywords and themes (Factor 1)

| **Reference** | **Themes and keywords** |
|---|---|
| Héroux-Vaillancourt et al. (2020) | research & development, research and development, r&d, researcher, product development, technology development, technical development, development phase, development program, development process, development project, development cent, development facility, technological development, development effort, development cycle, development research, development activity, fundamental research, basic research |
| The OECD Glossary of Statistical Terms | **Definition:** Research and development is a term covering three activities: basic research, applied research, and experimental development<br>**Cross References:**<br>**1-SNA**: Research and development by a market producer is an activity undertaken for the purpose of discovering or developing new products, including improved versions or qualities of existing products, or discovering or developing new or more efficient processes of production.<br>Research and development: a term covering three activities: basic research, applied research, and experimental development. Research and development services in natural sciences and engineering; social sciences and humanities and interdisciplinary.<br><br>**2- UNESCO:** Any systematic creative activity undertaken in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this knowledge to devise new applications. Includes fundamental research, applied research in such fields as agriculture, medicine, industrial chemistry, and experimental development work leading to new devices, products or processes. |

Table 2.4 R&D and relevant keywords and themes (Factor 1) (Cont'd and end)

| | |
|---|---|
| Collins Thesaurus of the English Language source: https://www.collinsdictionary.com/dictionary/english-thesaurus | **Research:** Sense of investigation**: i**nvestigation, study, inquiry, analysis, examination, probe, exploration, scrutiny, experimentation, delving, groundwork, fact-finding.<br><br>Sense of investigate: investigate, examine, experiment, explore, probe, analyse, look into, work over, scrutinize.<br><br>Sense of analyse examine, test, study, research, judge, estimate, survey, investigate, interpret, evaluate, inspect, work over.<br><br>Sense of analysis: examination, test, division, inquiry, investigation, resolution, interpretation, breakdown, scanning, separation, evaluation, scrutiny, sifting, anatomy, dissolution, dissection, assay, perusal, anatomization.<br><br>Sense of examination: inspection, testing, study, research, trial, checking, review, survey, investigation, analysis, consideration, observation, going-over(informal), vetting, scrutiny, appraisal, interrogation, assay, perusal, recce(slang).<br><br>**Development**: Sense of growth: growth, increase, growing, advance, progress, spread, expansion, extension, evolution, widening, blooming, maturing, unfolding, unravelling, burgeoning, advancement, progression, thickening, enlargement.<br><br>Sense of establishment: establishment, forming, generation, institution, invention, initiation, inauguration, instigation, origination,<br><br>Sense of event: event, change, happening, issue, result, situation, incident, circumstance, improvement, outcome, phenomenon, evolution, occurrence, upshot, turn of events, evolvement<br><br>Sense of advance: increase, rise, development, gain, growth, boost, addition, expansion, extension, enlargement, escalation, upsurge, upturn, increment, intensification, augmentation.<br><br>Sense of advancement: promotion, rise, gain, growth, advance, progress, improvement, betterment, preferment, amelioration<br><br>Sense of circumstance; detail, fact, event, particular, respect, factor. |

**Factor 2: Intellectual Property (IP)**

Intellectual property (IP) is related to mind creations, including patents, copyright, and trademarks, which could differ by country. The law gives people the right to profit and access information related to goods and inventions (Ambrose, 1990). This issue is designed for private and tradable property rights for different aspects of invention and innovation, and it creates private property rights around knowledge to invest resources in various activities such as designing and developing new products and techniques (Greenhalgh & Longland, 2001). Since innovation includes technological knowledge developments, technological innovations play a significant role in economic development (Baumol, 2002).Some definitions, such as "invention," refer to the "first idea, sketch or contrivance of a new-to-the-world product, process or system, which may or may not be patented"(Freeman & Soete, 1997, p. 201). However, invention and innovation are different from the concept of discovery (Granstrand, 1999). "Invention" refers to inventing something that does not exist and is invented by man; hence, when something is close to reproduction, copy or duplication of another thing, it is referred to as "imitation" (Granstrand, 1999). Based on OECD research, a patent is defined as "a right granted by a government to an inventor in exchange for the publication of the invention; it entitles the inventor to prevent any third party from using the invention in any way, for an agreed period".(O. F. Manual, 2013, p. 212). Patents, trade secret rights, and copyrights as IP and intellectual property rights (IPRs) have been recognized as innovations and technological developments since the 1930s. IP can be defined as one of the critical sources of competitive advantage at micro-level developments in many industries. Many studies show a connection between the importance of innovation and technological developments and economic growth and welfare (Baumol, 1986; Rosenberg, 1963; Scherer, 2011; Solow, 1956, 1957). Researchers also highlighted the relationship between technology and innovation management (Burns & Stalker, 1961; Chesbrough, 2003; Pavitt, 1990; Trott, 2008; Utterback, 1994). Some traditional disciplines were explained regarding the research related to the IPRs (Arrow, 1972; Romer, 2002). Research related to innovation IP management has been improving since the 1990s (Granstrand, 1999; Pisano & Teece, 2007; Reitzig, 2004). According to the concept of IP management and innovation, activities such as technology trade, licensing, collaboration research and development, crowdsourcing, acquisitions, and divestments are innovation activities. Table 2.5 presents different keywords, themes, and concepts related to IP from different resources.

Table 2.5 IP and relevant keywords and themes (Factor 2)

| **Reference** | **Keywords and themes** |
|---|---|
| Héroux-Vaillancourt et al. (2020) | Patent, intellectual property, trade secret, industrial design |
| The OECD Glossary of Statistical Terms | **Definition**: Intellectual property rights refers to the general term for the assignment of property rights through patents, copyrights and trademarks. These property rights allow the holder to exercise a monopoly on the use of the item for a specified period.<br><br>By restricting imitation and duplication, monopoly power is conferred, but the social costs of monopoly power may be offset by the social benefits of higher levels of creative activity encouraged by the monopoly earnings.<br><br>**Context**: Ownership of ideas, including literary and artistic works (protected by copyright), inventions (protected by patents), signs for distinguishing goods of an enterprise (protected by trademarks) and other elements of industrial property. |

Table 2.5 IP and relevant keywords and themes (Factor 2) (Cont'd and end)

| | |
|---|---|
| Collins Thesaurus of the English Language source: https://www.collinsdictionary.com/dictionary/english-thesaurus | **Intellectual: Sense** of mental**:** mental**,** cognitive, cerebral, physical.<br>Sense of scholarly: clever, intelligent, scholarly, learned, academic, lettered, rational, cerebral, erudite, scholastic, highbrow, well-read-studious, bookish,<br>Sense of academic: academic, expert, genius, thinker, master, brain(informal), mastermind, maestro, boffin(British, informal), highbrow, egghead, brainbox, pointy-head(informal, US), bluestocking(usually derogatory), egghead(informal), rocket scientist(informal), highbrow, master-hand, fundi( south Africa), acca( Australian, slang)<br>Sense of boffin: expert, authority, brain(s)(informal), intellectual, genius, guru, inventor, thinker, wizard, mastermind, intellect(informal), maven(US), fundi(South Africa)Sense of bookish: studious, learned, academic, intellectual, literary, scholarly, erudite, pedantic, well-read, donnish, swotty (British, informal).<br>Sense of brain: intellectual, genius, scholar, sage, pundit, thinker, master hand, intellect(informal), prodigy, highbrow, rocket science(informal), egghead(informal), brainbox, clever clogs, bluestocking(derogatory)<br>**Property:** Sense of possessions**:** possessions, goods, means, effects, holdings, capital, riches, resources, estate, assets, wealth, belongings, chattels,<br>Sense of land: land, holding, title, estate, freehold, realty, real property, Sense of quality: quality, feature, characteristic, mark, ability, attribute, virtue, trait, hallmark (British), peculiarity, idiosyncrasy<br>Sense of attribute: quality, point, mark, sign, note, feature, property, character, element, aspect, symbol, characteristic, indication, distinction, virtue, trait, hallmark, facet, quirk, peculiarity, idiosyncrasy<br>Sense of capital: money, funds, stock, investments, property, cash, finance, finances, financing, resources, assets, wealth, principal, means, wherewithal, wonga(slang)<br>Sense of characteristic: feature, mark, quality, property, attribute, faculty, trait, quirk, peculiarity, idiosyncrasy |

**Factor 3: Collaboration**

Rapid development in organizational complexity and global environments within the information and communication technology industries has caused some opportunists to study firms' behaviours in more depth. Close collaboration among firms and their partners is a result of a thriving innovation ecosystem. Based on research by Thomson, Perry, and Miller (2008), formal or informal interactions between organizations as a process of the reparative and iterative sequences of negotiation help to develop and evaluate commitments and execution of their work. The existing literature explains the reasons and motivations that encourage firms to collaborate for innovation purposes. Three main themes can be found for the benefits of collaboration: (1) sharing of resources and knowledge, (2) reduction of costs and risks, (3) improvements in performance and competitiveness. This categorization is covered particularly on the networks and business strategy (Gulati, Nohria, & Zaheer, 2000), and can occur at different levels of organizations (Bedwell et al., 2012). According to Geum, Lee, Yoon, and Park (2013) , there are four types of collaboration at the inter-organizational level: (1) competitors, (2) suppliers, (3) consumers, and (4) universities or research institutes. To decrease the risk of introducing innovation in a market, especially for new and complex products, collaboration with consumers and suppliers can be a good idea. “Collaboration” is working together to design and implement the best approaches for solving problems and delivering products that customers expect (Fawcett, Magnan, & McCarter, 2008). Information and communication technologies (ICT) are not exception from such collaboration changes. They are becoming increasingly important parts of an economy by facilitating the globalization of many services in different geographical locations. For more information on these concepts and the relevant keywords and themes, refer to Table 2.6.

Table 2.6 Collaboration and relevant keywords and themes (Factor 3)

| **Reference** | **Keywords and themes** |
|---|---|
| Héroux-Vaillancourt et al. (2020) | Affiliation, collaboration, cooperation, partners,Partnership, consorti, international consorti,Global consorti |
| The OECD Glossary of Statistical Terms | **Innovation co-operation**:<br>**Definition**: Innovation cooperation involves active participation in joint innovation projects with other organizations. These may either be other enterprises or non-commercial institutions. The partners need not derive immediate commercial benefit from the venture. Pure contracting out of work, where there is no active collaboration, is not regarded as cooperation. Co-operation is distinct from open information sources and acquisition of knowledge and technology in that all parties take an active part in the work.<br>**Context:** Innovation cooperation allows enterprises to access knowledge and technology that they would be unable to utilize on their own. There is also great potential for synergies in cooperation as partners learn from each other. |
| Collins Thesaurus of the English Language source: https://www.collinsdictionary.com/dictionary/english-thesaurus | Sense of teamwork: teamwork, partnership, cooperation, association, alliance, concert<br>Sense of conspiring: conspiring, cooperation, collusion, fraternization<br>Sense of alliance: union, league, association, agreement, marriage, connection, combination, coalition, treaty, partnership, federation, pact, compact, confederation, affinity, affiliation, confederacy, concordat<br>Sense of association: group, company, club, order, union, class, society, league, band, set, troop, pack, camp, collection, gathering, organization , circle, corporation, alliance, coalition, partnership, federation, bunch (informal), formation, faction, cluster, syndicate, congregation, batch, confederation, cooperative, fraternity(US, Canadian), affiliation, posse(slang), clique, confederacy, assemblage, social network<br>Sense of cooperation: teamwork, concert, unity, collaboration, give-and-take, combined effort, esprit de corps, concurrence, kotahitanga (New Zealand) |

**Factor 4: External Financing**[5]

This concept is related to the fund that a firm can obtain from other resources, such as outside investments. For example, Chen, Cheng, and Lo (2013) discussed that accounting restatements affect firms' private and public debt and equity. A focus on the 27 countries and their start-up financing entrepreneurial firms showed that entrepreneurs' experience in managing start-ups affects institutional investors (Nofsinger & Wang, 2011). Having significant financial commitment leads successful innovation requires. According to Miller and Bromiley (1990), it would be difficult to have innovation when financing is uncertain. Although innovation can be defined as a source of competitive advantages (Bates & Flynn, 1995), it can also damage firms, especially when R&D investment requirements are not acceptable (Dosi & Nelson, 1994). According to Barney (1991), a significant investment in innovation requires resources beyond the firms' internal cash; therefore, having access to external resources can increase competitive advantage. Evidence illustrates a relationship between financial constraints and firms' specific features, such as size and age (Hall, 2002). Firms can create an income stream via successful innovation mechanisms, and thus, financial resources can sustain firms' innovation potential (Brown, Fazzari, & Petersen, 2009). Studies declare a solid and positive relationship between firms' R&D investment and fixed asset sales funding (Borisova & Brown, 2013). According to Mulkay, Hall, and Mairesse (2001), this relationship can be more robust in different geographical situations with respect to R&D financing compared to the USA and France. There is a strong link between a firm's innovation and performance (Pillai & Rao, 1996). Moreover, the size of investments is vital for many R&D projects (Savignac, 2008). Table 2.7 presents keywords, themes and concepts related to the external financing from current study references.

[5] https://smallbusiness.chron.com/types-external-financing-80170.html

Table 2.7 External financing and relevant keywords and themes (Factor 4)

| **Reference** | **Keywords and themes** |
|---|---|
| Héroux-Vaillancourt et al. (2020) | Atlantic Canada Opportunities Agency, Business Development Bank of Canada, sustainable development technology, venture capital, Atlantic innovation fund, nrc-irap, fender, Industrial research, assistance program, grants, private investment |
| The OECD Glossary of Statistical Terms | Balance of payments, reserve assets<br>International reserve assets<br>**Definition:** Reserve assets consist of those external assets that are readily available to and controlled by a country's authorities for direct financing of international payments imbalances, for indirect regulation of the magnitude of such imbalances through intervention in foreign exchange markets to affect their currency's exchange rate, and for other purposes. The category of reserve assets defined in the IMF Balance of Payments Manuel, Fifth Edition comprises monetary gold, special drawing rights (SDRs), reserve position in the IMF, foreign exchange assets (consisting of currency, and deposits and securities), and other claims. |

Table 2.7 External financing and relevant keywords and themes (Factor 4) (Cont'd and end)

| | |
|---|---|
| Collins Thesaurus of the English Language source: https://www.collinsdictionary.com/dictionary/english-thesaurus | **External:** Sense of outer: outer, outside, surface, apparent, visible, outward, exterior, superficial, outermost<br>Sense of foreign: foreign, international, alien, exotic, exterior, extraneous, extrinsic<br>Sense of outside: outside, visiting, independent, extramural<br>Sense of alien: foreign, outside, strange, imported, overseas, unknown, exotic, unfamiliar, not native, not naturalized<br>Sense of apparent: seeming, supposed, alleged, outward, exterior, superficial, ostensible, specious<br>Sense pf exotic: foreign, alien, tropical, external, extraneous, naturalized, extrinsic, not native<br>**Financing:** Sense of funding: funding, money, support, funds, capital, subsidy, sponsorship, endowment, underwriting, financial support, financial backing, wonga(slang)<br>Sense of capital: money, funds, stock, investment(s), property, cash, finance, finances, financing, resources, assets, wealth, principal, means, wherewithal, wonga(slang)<br>Sense of endowment: provision, fund, funding, award, income, grant, gift, contribution, revenue, sunsidy, presentation, donation, legacy, hand-out- boon(archaic), bequest, stipend, bestwal, benefaction, largesse or largess, koha (New Zealand) Sense of money: cash, funds, capital, currency, wealth, hard cash, green( slang) readies(informal), riches, necessary(old-fashioned), silver, bread(slang), coin, tin(slang), brass( Northern England ,dialect), loot(informal), dough(slang), rhino( British slang, old-fashioned), the ready(informal), banknotes, dosh(British Australian , slang),lolly(British ,slang) the wherewithal, legal tender, megabucks(US, Canadian, slang) needful( informal), specie, shekels(informal), wonga (slang),dibs( slang),filthy lucre(facetious), moolah(slang), ackers( slang), gelt,(slang, US), spondulicks (slang, rare), pelf(archaic), mazuma ( slang, US), kembla( Australian, slang) |

**Factor 5: Creativity**

According to Drake (2003), creativity in firms is related to the collective or social process with additional insight into unique ways for economic activities, and has impact on the potentiality of global goods or services. According to Heunks (1998), there is a significant relationship between firms' success and their innovation and creativity attitudes; several factors, such as firms' growth, increasing productivity, profits, product innovation, process innovation, marketing innovation and R&D innovation, are discussed in this research. Companies always try to find new ways to ensure their innovation, and it is a significant concern to companies (Allen, 1984). Creativity factors were examined scarcely compared to the lack of knowledge about conditions that can enhance innovation performance. For example, one research focused on the firm's members, the impact of the team's communication was discovered as one factor that is important on creative performance (Leenders, Van Engelen, & Kratzer, 2003).

With respect to collect some keywords and themes related to creativity, Williams, Runco, and Berlow (2016), by focusing on 25-year creativity research in Web of Science articles, linked documents that have similar themes and analyzed all keywords by eight search strings related to creativity (e.g., creative process, creative product, creative style, etc.). Table 2.8 illustrates the creativity factor with its themes and keywords.

Table 2.8 Creativity and relevant keywords and themes (Factor 5)

| **Reference** | **Keywords and themes** |
|---|---|
| Williams et al. (2016) | innovation, work, management, organizations, perspective, performance, leadership,, teams, work environment, support, transformational leadership, contextual factors, product development, individual creativity, personality, intelligence, divergent thinking , divergent thinking, validity, reliability, factors, tests, artists, latent inhibition, scale, implicit theories, achievement, young children, openness performance, idea generation , self-efficacy, productivity loss, brainstorming groups, employee creativity, thought, job satisfaction, leader-member exchange, social innovation, goal orientation, psychological empowerment, strategies, activation students, education students, school, community, representation, teachers, program, teaching, curriculum, teaching/learning strategies, design process creative process, anterior cingulate cortex, working memory, labor, modulation, poetry, metaphor problem solving, insight, critique, youth, adolescence, stimulus-independent thought, hierarchy, task, depression, identity, working-memory capacity, dreams, incubation, chance discovery, challenge drawing, human-computer interaction, methodology, outcomes, craft, leaning, examples, constraints, teaching communication, evolution, communication, creation, adaptation, dementia, computational creativity, process, brain, chaos, interaction, psychoanalysis, systems, planning, agencies, frontotemporal dementia , creative problem solving, perception collaboration, war, play, way, creative economy, ethics, resilience, evaluation, emergence, school-children, stress, creative industries, perception, growth, engineering design, mood ,memory, information, positive affect, selective retention, implicit, blind variation, music, systems, hedonic tone, intrinsic motivation, flow, autonomy, experience, similarity, user involvement, patterns, regeneration, dimensions, literature, pedagogy ,creative potential, knowledge, aesthetics, aesthetics, transference, context, genius, media, fixation, schizophrenia, behavior, creative writing, experiences, magical ideation, future, risk, writers ,science , constructivism, history, discourse, love, philosophy, transformation, space, gender, styles, brain plasticity, prevention, classroom , entrepreneurship, mediation, imagination |

Table 2.8 Creativity and relevant keywords and themes (Factor 5) (Cont'd and end)

| | |
|---|---|
| Collins Thesaurus of the English Language source: https://www.collinsdictionary.com/dictionary/english-thesaurus | Sense of imagination: imagination, talent, inspiration, productivity, fertility, ingenuity, originality, inventiveness, cleverness, fecundity, imaginativeness.<br>Sense of cleverness: dexterity, ability, talent, gift, flair, ingenuity, adroitness<br>Sense of fecundity: fertility, creativity, inventiveness, fruitfulness, productiveness, fructiferous<br>Sense of ingenuity: originality, genius, inventiveness, skill, faculty, flair, knack, sharpness, cleverness, resourcefulness, shrewdness, adroitness, ingeniousness |

# CHAPTER 3 METHODOLOGY

This chapter consists of four main sections. Section 3.1 starts with problem statements and research questions. Section 3.2 covers research objectives. The research framework is illustrated in section 3.3, and finally, information regarding the data and methodology is presented in section 3.4.

## 3.1 Problem statement

A group of people, organizations, or objects collaborating or competing or linked by other means or for other purposes can be called as social networks (Wasserman & Galaskiewicz, 1994). Twitter usage in the industry as one of the popular social networks' research has been selected in different research domains such as the emotional behaviours analyzing, the user's political opinion mining or the influencers' behaviours on buyer decisions. To achieve these goals, using some tools like sentiment analysis can help researchers to collect more information about their industries (Anjaria & Guddeti, 2014; Bae & Lee, 2012; Bakshi, Kaur, Kaur, & Kaur, 2016; Bakshy, Hofman, Mason, & Watts, 2011; Eysenbach, 2011; Kontopoulos, Berberidis, Dergiades, & Bassiliades, 2013; Minh, 2013; Pak & Paroubek, 2010; Stieglitz & Dang-Xuan, 2012; H. Wang, Can, Kazemzadeh, Bar, & Narayanan, 2012; Weller et al., 2014).

Despite the great interest in analyzing social media aspects, few studies have analyzed Twitter text, especially with regards to the generation of content related to innovation concepts among companies' online conversations. Besides, some technological and business-driven solutions are not going to be enough. To address these research gaps and advance the growing literature on analyzing tweets based on innovation concepts, researchers investigated firms' Twitter accounts via a study on a particular case of the industrial subsector related to the Custom Computer Programming Services (NAICS code 514 511). The research questions are as follows:

RQ1 -To what extent are firms interested in innovation in their tweets?

RQ2 - What are the most frequent words used in the tweets of firms in that sector?

RQ3 - To what extent does LDA effectively identify topics related to innovation in the firm's tweets?

In this study, we use the firms' Twitter account information as a regular user to answer the above research questions.

## 3.2 Research objectives and contributions

This research project aims to demonstrate that insight can be extracted from the collected Twitter data using the proper tools, such as data mining, NLP algorithms, and web scraping. By focusing on innovation concepts, it would be easy to analyze tweets' data regarding different innovation topics.

Hence, our present study aims to:

OBJ1 – Collect Twitter data using web-scraping from companies of the Custom Computer Programming Services (NAICS code 514 511).

OBJ2 – Prepare (pre-processing) the Twitter data for the exploratory analysis.

OBJ3 – Exploratory data analyses on the collected data to identify the most important topics mentioned in tweets

OBJ4 – Identify whether these topics are related to innovation.

## 3.3 Research framework

This research is related to analyzing tweets. Due to the relative novelty of social network research studies for data collection and analysis, it applies ML and NLP to appropriately address the research questions. Figure 3.1 illustrates the research framework.

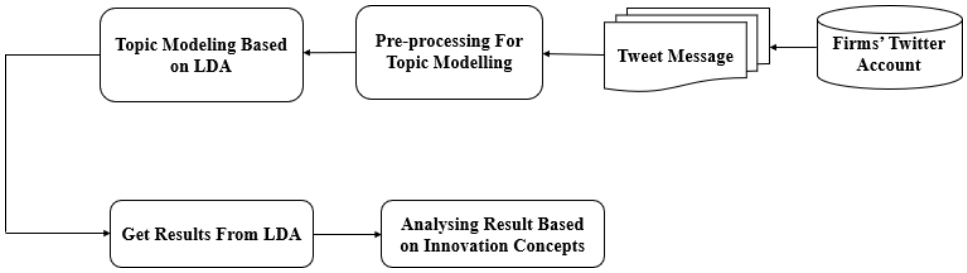


Figure 3.1 Research framework

## 3.4 Research data

This research focuses on Twitter accounts of Custom Computer Programming Services sectors (NAICS[6] 541511). This sector is further subdivided into two subsectors: 1) software publisher (511210) and 2) computer system design services (514512).

Using a secondary data source collected by the company Aberdeen Technology Data Cloud[7] in an international survey, information related to firms like their names, homepage links, and provinces were collected. Aberdeen Technology Data Cloud (ATDC)[8] collects data across all industries and company sizes with 20 technology areas ranging, and in 2017, they centred business intelligence content on five topic areas (1) hardware (2) software, (3) storage, (4) networking and (5) telecom. This group, by using some trained research assistance, tries to collect and update information.

**An overview of the NAICS codes 514512 and 511210**

These days, few companies or organizations can be found that have nothing to do with software applications, network equipment, and information technology. Companies who work in these areas always face different challenges related to the competition, market selling and customer services, etc. By developing, implementing, supporting, and managing systems, these companies like to have better, faster and safe relationships and connections with their audiences. The use of social networks to maintain relationships with others provides better business communication opportunities. Investigating scientific research to identifying companies' challenges can bring organizations closer to overcoming their obstacles. In this research, we focus on the Twitter

---

[6] It is a standard for the classifying business establishments based on the similarity in the processes in the products or services and each industry will have at least one associated NAICS code. More information can be appeared on appendix A.

[7] Data collection is carried out by research assistants from the Aberdeen Technology Data Cloud (ATDC) group, who interview companies in each sector on a monthly basis to keep their data collection up to date. This data set content across 20 technology areas and different business locations or sites with five topic areas: 1-Hardware – Servers, PCs and Printers, 2-Software – DBMS, ERP/CRM, Disaster Recovery, Security 3- Storage – Both the storage devices and the associated management software 4- Networking – Network LAN equipment and IP-centric software and services and finally 5- Telecom – Both voice and data platforms and services.

[8] https://www.aberdeen.com/research/#whyaberdeenresearch

accounts of customer computer programming firms. Information about these firms, such as location and homepage, are provided in the Aberdeen data set.

Since this research is focused on these firms' Twitter accounts, the recommended first step is to follow the homepage link to visit the firms' websites and become familiar with their work. The domain of the firms is computer services. Focusing on the Aberdeen data set individually for those codes shows that 56% of the computer system design services firms (code 514512) have submitted their webpage link in the Aberdeen data set. For software publishers (code 511210), this figure is 46%. Figure 3.2 illustrates these percentages.

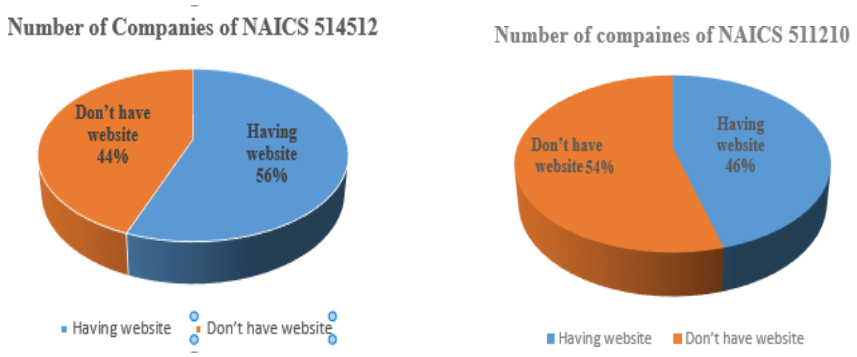


Figure 3.2 Information of the percentage of having home page in the dataset for codes NAICS 514512 and 511210

Visiting the 2,900 homepage links in the Aberdeen data set one by one revealed that 1,246 units have active homepage links with information about the firms. Figure 3.3 presents this information.

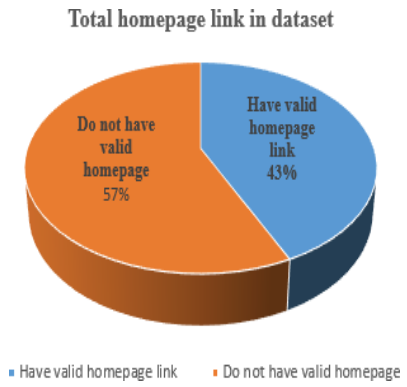


Figure 3.3 Total homepage link in the dataset for codes NAICS 514512 and 511210

By observing these 1,246 websites and focusing on the presence of icons or links related to social media, we found that 518 companies have social media icons or links on their homepage. Narrowing in on the objective of our research, 405 firms have activities in their Twitter account, and 162 have inactive (closed or not found) Twitter icons or links in their pages. Figure 3.4 presents the percentage of firms in these subsectors actively using Twitter.

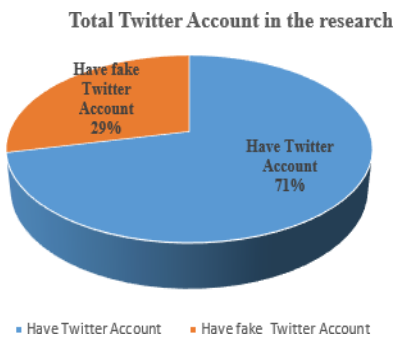


Figure 3.4 Percentages of using the Twitter account for codes NAICS 514512 and 511210

**Type of data**

Focusing solely on the 405 firms which have active Twitter accounts, the first raw data was extracted for a two-week duration by using web scraping by python codes and their libraries to collect specific fields (variables or features). The raw data was sorted in JSON files. The data collected as a result of this extraction fills 25,989 rows and 12 columns. Table 3.1 shows all variables in this data set.

Table 3.1 Data set variables

| N | Twitter Feature name | Types of data | Definition in the data set |
|---|---|---|---|
| 1 | Company | Object | Name of the company |
| 2 | tweet_usename | Float 64 | Username of the company |
| 3 | tweet_date | Object | Date of the tweet |
| 4 | tweet_retweets | Int64 | Retweeting the tweet |
| 5 | tweet favorites | Int 64 | Being favorites by others |
| 6 | tweet text, | Object | Context of the tweet |
| 7 | tweet_geo_location | Float64 | Location of the firm while tweets |
| 8 | tweet_mentions | Object | Reference usernames in the tweet text by using "@" sign |
| 9 | tweet_hashtags | Object | Metadata tag by the user by "#" sign |
| 10 | tweet_id | Int64 | Especial number for the tweet |
| 11 | tweet_permalink | Object | Link user put in the tweet |
| 12 | tweet_text_comp | Object | Time and date text by company |

**Preparing data for topic modelling**

Since one of objective of this study is focusing on using text mining and the LDA algorithm to find topics in firms' Twitter accounts, the following steps were needed to implement this methodology:

**Step 1: Reading data**

In this step, we collected users' Twitter data and saved each company's information as one JSON file using Python web scraping. All 405 JSON files were converted into one CSV file. Using Python code helped us to know more about data set size. The size of this data set is 25,989 rows of tweets, with 12 columns of different variables with different types of data: Company name, tweet_usename, tweet_date, tweet_retweets, tweet _favourites, tweet _text, tweet_geo_location, tweet_mentions, tweet _hashtags, tweet_id, tweet_permalink and tweet_text_comp with different types of data. For this dataset, 2.4+ MB memory is allocated. Regard to using LDA topic modelling and analyzing, 510,979 tweet_text is collected.

**Step 2: Data pre-processing** [9]

Concerning using text mining techniques and topic modelling from unstructured data, several preparation steps are required. Topic modelling, one of many automated statistical methods, can help to discover topics within text content. Pre-processing is a crucial stage of topic modelling and generally involves the following steps: (1) removing punctuation, (2) tokenization for breaking data into single words, (3) removing stopping words (eliminating words like “and,” “or,” “what”), (4) stemming (removing suffixes and prefixes), and (5) word lemmatization to transform different word forms to their simplest expression (e.g., sing, sang, sung, singing). Python’s[10] NLTK library was used for performing lemmatization, stemming, and stop word processes

**Lemmatization**: This is used for word morphological analysis to remove inflectional ending to return the dictionary form of a word. For instance, the term “seeing” is changed to “saw” or “see.”

**Stemming** This refers to removing affixes. For instance," studying" is changed to "study".

**Tokenization**: This is the task of dividing documents into pieces called tokens. Cleaning punctuation and adding stop words: using libraries “English” in NTLK helped clean some not significant words.

**Scikit-learns CountVectorizer for Feature Extraction**: The use of libraries (“English” in NTLK) helped clean some insignificant words.

**Topic modelling using LDA**

The process of applying statistical models (topic models) to extract the hidden (latent) topics in the text data is called topic modelling. Discovering hidden patterns in the text can include several steps: (1) Latent Semantic Analysis (LSA), (2) Probabilistic Latent Semantic Analysis (PLSA), (3) Latent Dirichlet Allocation (LDA), (4) Correlated Topic Model (CTM), (5) Explicit Semantic Analysis, (6) Hierarchical Dirichlet Process, and (7) Non-negative Matrix Factorization (Negara, Triadi, & Andryani, 2019). Several applications can be named for topic modelling. For instance, regarding Boyd-Graber, Hu, and Mimno (2017), information retrieval (IR), smoothing language models,

---

[9] Codes related to the implementation are located in appendix B.

[10] Scikit-learn in Python 3.7 helped for implementing the model.

query expansion, and topic models can search for personalization. As another application of topic modelling, topical changes can be used in various fields such as newspapers, historical records, and historical scholarly journals. This concept can also be used in the literary world to analyze authors' emotions, thoughts, and fictional cheaters. Using topic modelling in online discussions across social media platforms as another application of topic modelling can help researchers understand the impact social media has on people's behaviours, such as using companies' products or participating in political voting events.

Since topic modelling is a kind of unsupervised algorithm, using this model helps to understand and extract hidden topics in the text. Several topic modelling techniques are available, for instance, LDA, Biterm and clustering word embedding (Jónsson & Stolee, 2015).

In this chapter, we employ the LDA as proposed by David M Blei et al. (2003b).

To avoid confusion, some terms related to this model need to be described.

**Word:** This is an item of the vocabulary {1,..,V} in NLP; unit basis vectors can be presented by words and one component equal to one and others equal to zero.

**Document** If N is the total of the words in the collection, and wn is the nth word of the N, the collection of the words is documented and can be illustrated as $D=\{w_1,..,w_N\}$.

**Corpus:** If M is the total number of documents, a collection of the documents is a corpus, which can be shown as $C= \{D_1, \dots , D_M\}$.

**Term/Token** These are the building blocks of documents, which include words, phrases, symbols, or any meaningful element in a document (Wedenberg & Sjöberg, 2014).

**Bag of words** Each document is represented as a collection of individual words in which some issues, such as grammar, word order, sentence structure, and punctuation, are ignored.

**Vectorization**: This is a methodology for mapping words or phrase to vector.

**Tf-idf:** This is a statistical method to determine the degree of importance a word has in a document in a collection for analyzing textual data. The number of times a term appears in a document is the term frequency, and the logarithm of the number of documents in the corpus divided by the number of documents is called the inverse document frequency.

t= term d=document D= all documents then; tfidf (t, d, D)=t f (t, D) and

t f (t, d) = frequency (t, d) idf (t, D)=log $N/|\{d \in D : t \in d\}$

The algorithm is based on a probabilistic model that should yield interpretable topics within a corpus. As such, performing LDA in large documents has the ability to summarize and cluster words in the document (J. C. Campbell, Hindle, & Stroulia, 2015), and allocate them into coherent topics. The topics – word classification into topics – are, in essence, latent (hidden) while the documents are observable. The LDA model assumes that a document generally comprises few topics and that a few words can characterize a topic. The underlying strategy of the LDA model first postulates the existence of K topics specific to a collection of documents, and then the model seeks to find the mixture of topics that make up each document. To do this, the algorithm will iterate by successively assigning a weight to each word for each topic, then a weight for each topic to each document. The mixture of topics for each document is drawn from the Dirichlet probability distribution. In this project, we load the LDA model from SK-learn[11].

According to Negara et al. (2019), the procedure of this algorithm is as follows:

"...***Procedure****:*

***Input****: Number of document M*

*Number of topics t*

*β Vocabulary matrix*

***Output****: Topic probability distribution for each word in document.*

***Steps****:*

*1. Choose the topic distribution α*

*2. Assign each word W in a document d to one of the t topics.*

*3. For each word W in a document d*

*• For each topic calculate P (Topic t | Document d)*

---

[11] Scikit-learn (known as SK-learn) is a Python free software machine learning library with different algorithms (Pedregosa et al., 2011).

• *Calculate P (word W | Topic t)*

*4. The selection word W for a topic t depends on the distribution of β vocabulary words"* Source : (Negara et al., 2019)

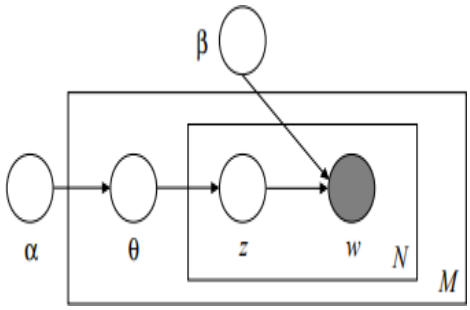


Figure 3.5 A LDA graphical model source : (David M Blei et al., 2003)

LDA is a three-level hierarchical Bayesian model for discrete data collection and documents with their words , represent randomly over latent topics, and the distribution over the words can result in a latent topic (David M Blei, Ng, & Jordan, 2003a). Based on the graphical model of LDA in Figure 3.5, boxes represent replicates. Documents are represented as the outer circle (plate) and repeated choice of topics. Words within a document can be represented by the inner circle (plate).

One of the concepts related to the LDA topic modelling is hyper-parameters. Hyper-parameters are input into any ML model that generates its own parameters to influence the values of said generated parameters in the hope of making the model accurate"[12]. Table 3.2 Lists the hyper parameters of the LDA.

[12] https://towardsdatascience.com/hyper-parameter-tuning-and-model-selection-like-a-movie-star-a884b8ee8d68

Table 3.2 LDA hyperparameters

| **Parameter Name** | **Description** |
|---|---|
| Num_topics | Number of topics for LDA to find within the data and it required to use positive integer to set |
| Feature_dim | It is a positive integer and show the Size of vocabulary of the input document corpus |
| Min _batch_size | It is positive integer which shows the total number of documents in the input document corpus |
| Alpha | It is optional and positive float valuable .it shows the sum of the element of the Dirichlet prior |
| Max_restarts | It is optional positive integer which helps to find better quality performance |
| Max_interations | It is optional positive integer for finding better quality computation |
| tol | It is optional positive float related to the target error tolerance. |

**Log-likelihood**

According to Nallapati and Cohen (2008), to measure how well the models predict unobserved data, log-likelihood can be applied. In the model, the higher log-likelihood, the better the model is at predicting for unseen data.

$$L = (p, m. W) = \sum_{i=1}^{k} \sum_{j=1}^{N} (logp - \frac{d}{2} log2\pi - \frac{1}{2} \log |W| - \frac{1}{2} (x_{ij} - m_i) w^{-1} (x_{ij} - m_i)^t)$$

**Perplexity**

According to David M Blei et al. (2003b), using perplexity can help to evaluate the models . A lower perplexity score presents the better generalization performance. According to Mirylenka, Scotton, Miksovic, and Dillon (2019), the number of latent topics is a user-defined parameter. Some measurements, such as perplexity, can help to know how well the probability distribution by model predicts, and it calculates as follows:

$$Perplexity = exp^{-\frac{1}{n} \sum_{i=1}^{n} \ln P(a)}$$

**Visualization of LDA topic modeling**

In order to show the simple probability distribution of words, the PyLDAvis library can be used. It helps to present topics and their relative and similar corpus among topics. Salience and relevance are two measures that can be adjusted interactively (Sievert & Shirley, 2014). According to Blanco, Pérez-López, Fdez-Riverola, and Lourenço (2020), one of the number of topics is related to the hyper-parametrization of the LDA. Chuang, Manning, and Heer (2012) introduced saliency to explain the terms that best discriminate topics concerning other topics.

Based on Blanco et al. (2020), saliency can be defined as follows:

"*… and term saliency described how informative the term w was for determining the generating topic, versus a randomly selected term in an information-theoretic sense. Distinctiveness was evaluated based on the Kullback and Leibler (1951) divergence between the conditional probability P(T w), i.e., the likelihood that the observed term w was generated by the latent topic T, and the marginal probability P(T), i.e., the likelihood that any randomly selected term w was generated by topic T*

$$\text{distinctiveness}(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$

$saillanc\,(w) = P(w) \times$ distinctiveness". Source: (Blanco et al., 2020).

According to Chuang et al. (2012),relevance measurement in PyLDAvis is used to show the ranking of the terms within a topic.

P(w)= the marginal Probability of w

P(w|T) = probability of term w given by topic T

Relevance $(w, T|\lambda) = \lambda log$ (w|T) + $(1 - \lambda)$ $log$ ( $P$(w|T) / $P(w)$ )

$\lambda$= hyperparameter $0<\lambda<1$ using for adjusting the list of terms of topics

$\lambda = 1$ select words based on hierarchization classical method

$\lambda = 0$ select the words based on the probability distribution of the word in the topics and the corpus. If the word in the corpus is rare, it is more likely to be chosen by the algorithm.

# CHAPTER 4 RESULTS

This chapter reports results and findings of research based on implementing LDA topic modelling and text mining. Section 4.1 presents innovation factors that are used in this chapter and the results of LDA topic modelling. Section 4.2 presents discussions about these results.

## 4.1 Innovation concepts and LDA topic modelling results

To understand the innovation concepts and analyze and compare LDA topic modelling results with these concepts, several resources are used in this research. First of all, in accordance with Héroux-Vaillancourt et al. (2020), we collected the following four factors and keywords related to them: R&D, collaboration, external financing, and intellectual property (IP). The second resource is the research by Williams et al. (2016) that is related to creativity and its relevant keywords. And finally, in striving to achieve the best results, we used other resources such as the older Oslo Manual OECD (2005) , O. Manual (2018) and glossary of the statistical terms web page by OECD [13]. Since Twitter users can write informally, the Collins dictionary helped us compare and find results regarding the use of slang and informal words in tweets.

**Results and findings by text mining and LDA topic modelling**

Since the first research question of this study is related to the firms' interest in using innovation concepts in their tweets, text mining via LINQ[14] technology in C#[15] codes in the Visual Studio helped us to count keywords related to the four innovation factors in our sample of tweets. In fact, in the first step , focusing on the innovation factors (R&D, IP, collaboration and external financing) collected from the Héroux-Vaillancourt et al. (2020) helped to point out the view of using these innovation factors in Tweets. Table 4.1 and tweet exploration by text mining show that firms are more interested in using collaboration and external financing concepts and less interested in using IP concepts in their tweets.

---

[13] The OECD Glossary of Statistical Terms

[14] LINQ (Language Integrated Query) is uniform query syntax in C# and VB.NET to retrieve data from different sources and formats source: https://www.tutorialsteacher.com/

[15] C# is a new language created by Microsoft and submitted to the ECMA for standardization source : https://www.developer.com/

Table 4.1 Number of times of using innovation factors and their keywords source: Héroux-Vaillancourt et al. (2020)

| **Factors** | **Times of using these factors and keywords and themes related to them** |
|---|---|
| R&D | 109 |
| IP | 16 |
| Collaboration | 576 |
| External financing | 227 |

The researcher has added creativity as a fifth innovation concept to expand the exploration of tweets based on innovation factors in Table 4.1. Moreover, using other resources such as Oslo Manual OECD (2005) , O. Manual (2018) and glossary of the statistical terms web page by OECD, Williams et al. (2016) and Collins dictionary helped us develop deep and wide ranges of keywords, themes, and concepts for all five current research factors.

To evaluate the first research question's findings and answer the second research question, which is related to the most frequent words in tweets, implementing LDA topic modelling and using pyLDAvis[16] library in python, can facilitate visualization of results[17]. Figure 4.1 presents the 27 most salient terms of the data set. In this figure, bubble size is directly related to topic prevalence. A big non-overlapping bubble represents good topic modelling; overlapping can appear when there are similar terms or many topics. Overall term frequency is illustrated by the colour blue, and estimated term frequency based on each topic is illustrated by the colour red.

[16]https://pypi.org/project/pyLDAvis/#:~:text=**pyLDAvis**%20is%20designed,an%20interactive%20web%2Dbased%20visualization.

[17] https://github.com/bmabey/pyLDAvis

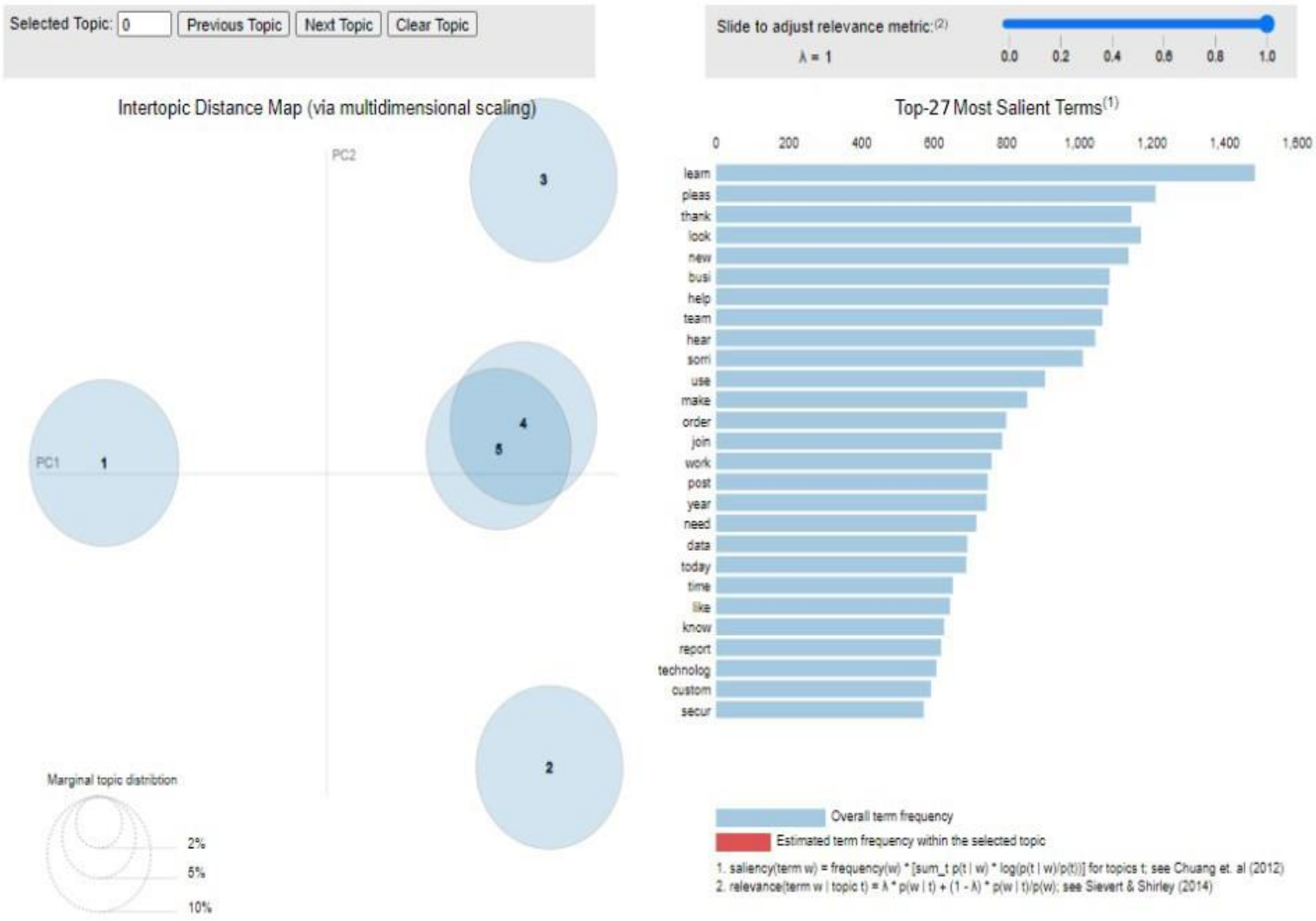


Figure 4.1 The 27 most salient terms in data set

Figure 4.1 shows an overlap between topics number four and five, demonstrating that some of the concepts of these two topics are similar. For example, keywords related to market activities can be seen in both topics[18]. The right side of Figure 4.1 shows the 27 salient terms of the data set that are most repeated.

With regard to focusing on the five innovation factors and answering the third research question, which focuses on identifying related topics, clicking on each bubble shows the relevant keywords for each topic. Comparing each topic and its words (see Figures 4.2 to 4.6) with research reference tables can guide us to categorize innovation factors in tweets. These findings are explained below.

**Analysing topic one based on five innovation concepts**

Figure 4.2 illustrates the words and themes related to topic one. According to this figure and its word, the stemming and lemmatization pre-processing steps caused several changes in some words, resulting in the appearance of incomplete words. These words are “industri,” “sorri” and “chang.”. The use of reference tables and different resources revealed some replacement words based on the root of each word. For example, "sorri" can have a different meanings: (1) a conjugation of a second person or the singular imperative of the verb “sorrir”[19] in Asturian and Portuguese, (2) based on communications language in tweets, it can be “sorry.” Another word in this list is “industri,” which could refer to “industrious,” “industrial,” “industrialization,” and other related words. Finally, the word “chang” could refer to change,” “changeable,” “changing,” etc.[20]

Focusing on topic number one (Figure 4.2) and its relevant words, the findings show that most of the keywords in this figure are related to creativity (11 words). In contrast, IP and external financing concepts are less used. The number of keywords and themes related to collaboration and R&D are the same (5 words). Table 4.2 illustrates and categorizes words based on the five innovation factors.

---

[18] Based on the aim of the research, different kinds of topics can be present by the same results.

[19] See: https://en.wiktionary.org/wiki/sorri

[20] See : https://www.vocabulary.com/dictionary or https://www.etymonline.com

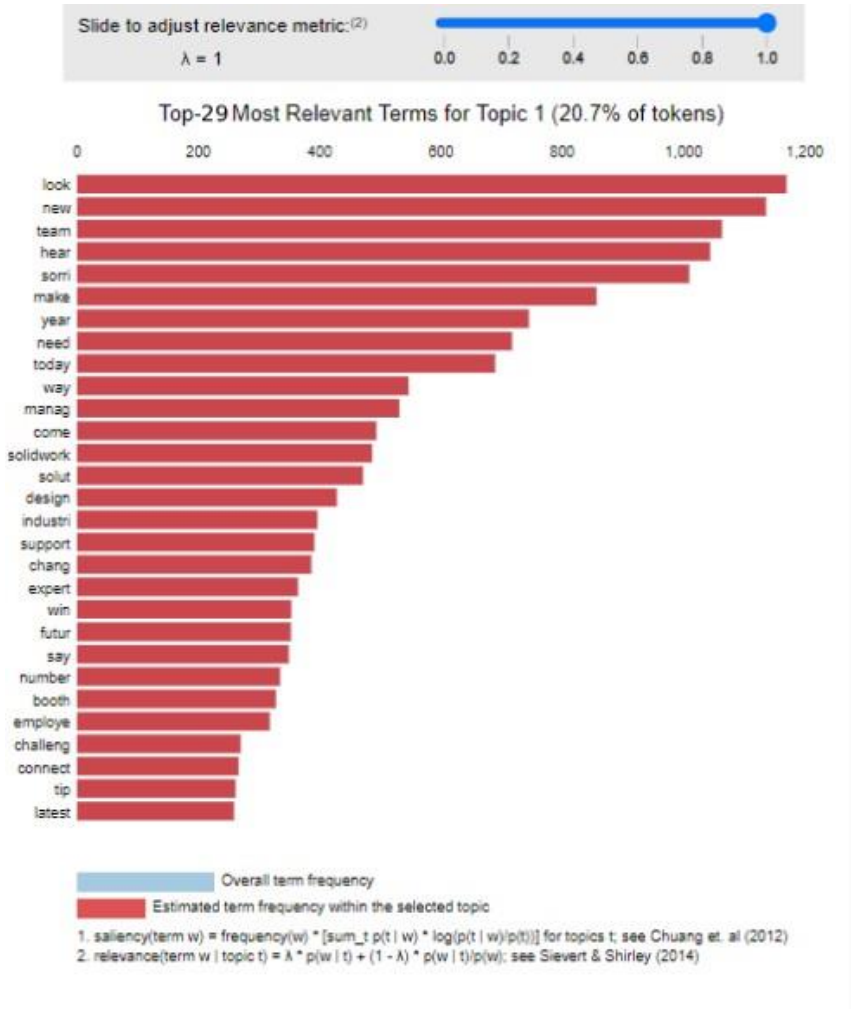


Figure 4.2 Relevant terms for topic one

Table 4.2 Results of topic one by focusing on innovation factors

| **Factors** | **keywords** | **Probable topic** |
|---|---|---|
| R&D | Look-new-industri-chang-expert | Creativity |
| IP | Design -expert | |
| Collaboration | New-team- need-support-connect | |
| External financing | Need-Support-tip | |
| Creativity | New-team-make-way-manag-design-support-change-futur-employ-challeng | |

**Analysing topic two based on five innovation concepts**

According to the results shown in Figure 4.3, some words are not in complete dictation. These kinds of results are mainly because of using stemming and lemmatization steps at the pre-processing level. For instance, “creat” can be replaced by “create,” “creating,” and “created” or similar words with the same root. Additionally, “innov,” can be “innovating,” “innovation,” “innovative,” and “innovational,”etc.[21]

Focusing on topic two words (Figure 4.3) and comparing the findings with the five innovation research factors, the LDA results show that terms related to creativity have the most frequency in tweets (5 words). In contrast, the IP theme is not present. External financing is the second factor that is used less in this topic. Keywords related to R&D and collaboration appear in the same number (4 words). Hence, the probable topic for topic number two would be creativity. Table 4.3 categorized findings based on research innovation factors and results.

Table 4.3 Results of topic two by focusing on innovation factors

| **Factors** | **Keywords** | **Probable topic** |
|---|---|---|
| R&D | Technology-creat-service-program | Creativity |
| IP | | |
| Collaboration | Open-innov-technology-share | |
| External financing | Technology-Service | |
| Creativity | Program-job-plan-open-innov | |

[21] See : https://www.vocabulary.com/dictionary or https://www.etymonline.com/

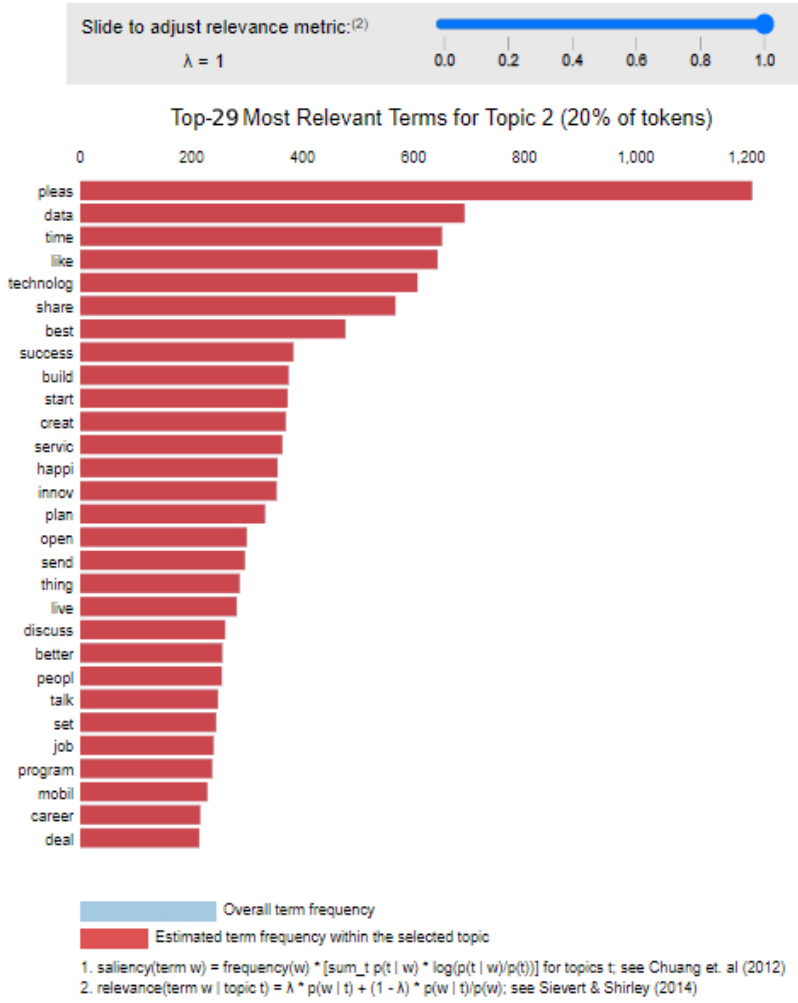


Figure 4.3 Most relevant terms for topic two

**Analysing topic three based on five innovation concepts**

Figure 4.4 shows the words related to topic three. As we can see regarding the using pre-processing steps, in the results, the word "opportun" could be "opportunity" or "opportune".[22] According to this finding, keywords related to creativity (6 words) are most used in this topic (see Table 4.4). Collaboration and R&D with four words in the list is the second topic; however,

[22] See : https://www.vocabulary.com/dictionary or https://www.etymonline.com/

keywords related to IP and external financing appear less in this topic. Therefore, creativity is a probable topic for topic number three.

Table 4.4 Results of topic three by focusing on innovation factors

| **Factors** | **Keywords** | **Probable topic** |
|---|---|---|
| R&D | Use-work-know-read | Creativity |
| IP | | |
| Collaboration | Work-know-partner-organ | |
| External financing | Opportun | |
| Creativity | Leader-transform-organ-know-work-use | |

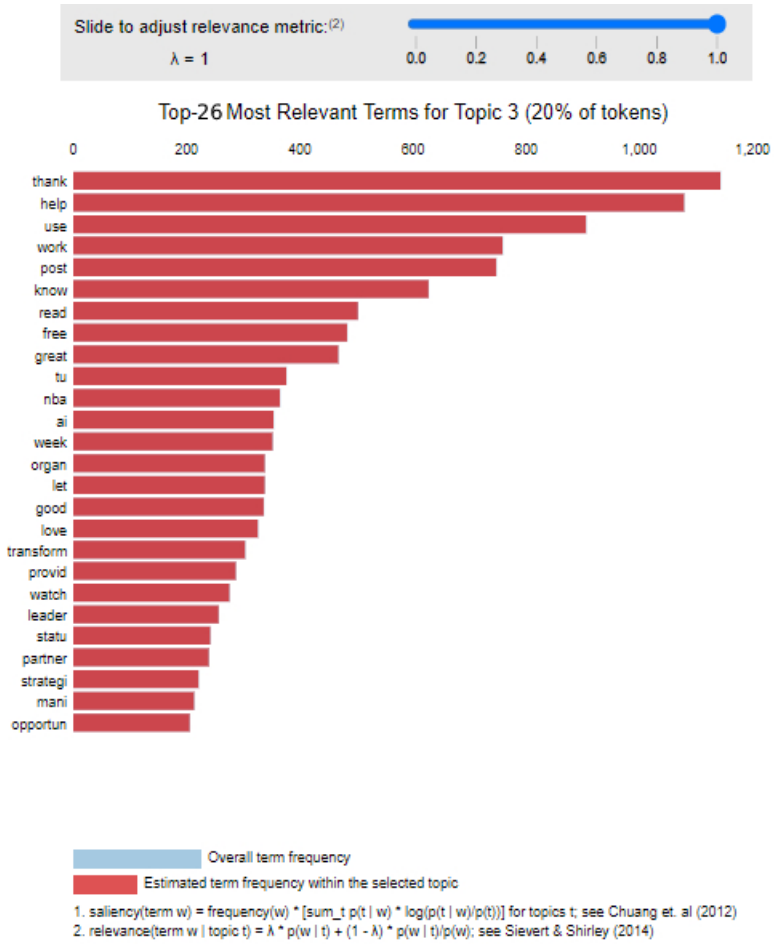


Figure 4.4 Most relevant terms for topic three

**Analysing topic four based on five innovation concepts**

According to the results shown in Figure 4.5, some words such as “experi,” “readi,” “secur,” and “busi” have changed due to the pre-processing steps, and they can be different words [23]. For example, “expri” can be replaced by “experiment,” “experience,” “experimental” or other relevant words. Or “readi” can be “readies,” which is one of the informal words for money[24], or it could be “reading” as well. “Secure” and “security” can be replaced for “secur,” and finally, “business,” “businessman/businesswoman,” “businesslike” or “business development” can be used instead of "busi"[25]

According to the LDA results on topic number four (Figure 4.5), shown in Table 4.5, the number of keywords related to external financing are the five words that are the most compared to other factors. With four words, collaboration, R&D, and creativity are probable second topics. IP keywords are less used in this topic. Therefore, external financing as an innovation factor can be a possible topic for topic number four.

Table 4.5 Results of topic four by focusing on innovation factors

| **Factors** | **Keywords** | **Probable topic** |
|---|---|---|
| R&D | Order-check-experi-tech | External financing |
| IP | Trade-right | |
| Collaboration | Order-network-tech-assist | |
| External financing | Assist-readi-right-secur-busi | |
| Creativity | Experi-power-leader-train | |

[23] See : https://www.vocabulary.com/dictionary or https://www.etymonline.com/

[24] See: https://www.collinsdictionary.com)

[25] It also can be used instead of bisy and bisy itself refers to“bysy, bisie, bysie, bisi, bysi, bysye, bisiȝ, besi, besy, besie, besye, busy, busi, busie” source : https://en.wiktionary.org/wiki/bisy#Middle_English

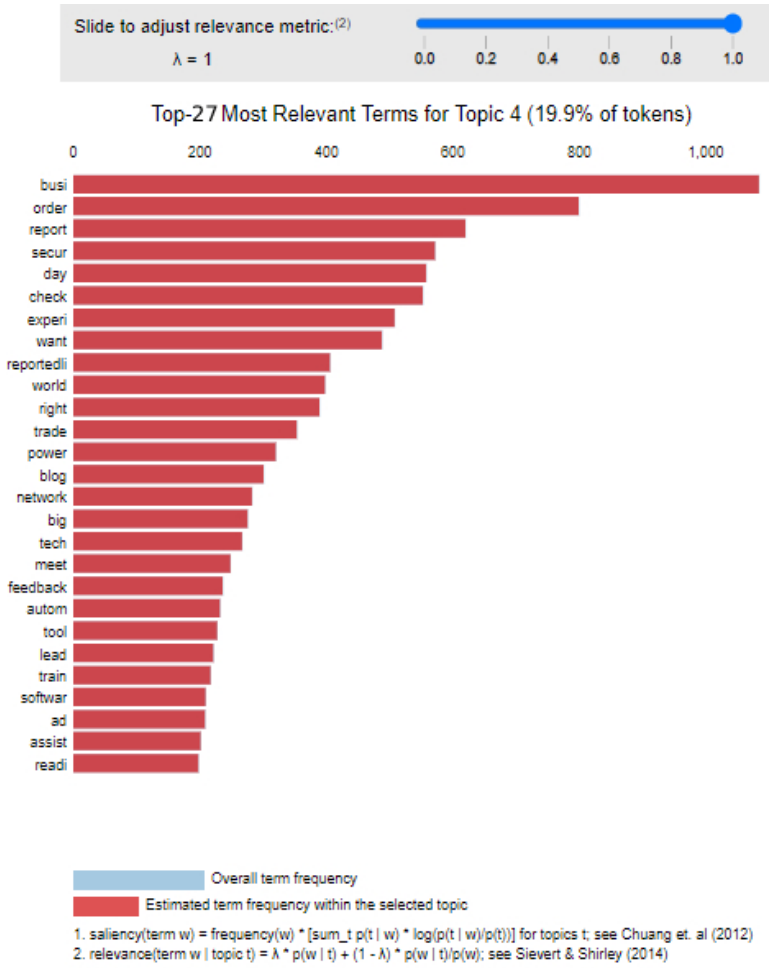


Figure 4.5 Most relevant terms for topic four

**Analysing topic five based on five innovation concepts**

According to topic number five (Figure 4.6), shown in Table 4.6, pre-processing steps made some changes in the word "compani," which can be "companion," "companionship" and other relevant words with the same root[26] .

[26] See: https://www.vocabulary.com/dictionary or https://www.etymonline.com

Table 4.6 shows that creativity keywords and themes are most used in this topic (7 words). R&D, with six words, can be the second topic, and collaboration, with two words, the third, whereas IP and external financing are less used in these findings.

Table 4.6 Results of topic five by focusing on innovation factors

| **Factors** | **Keywords** | **Probable topic** |
|---|---|---|
| R&D | Market-app-product-develop-improv-process | Creativity |
| IP | Feature | |
| Collaboration | Join-compani | |
| External financing | Market | |
| Creativity | Learn-product-develop-think-process-event-insight | |

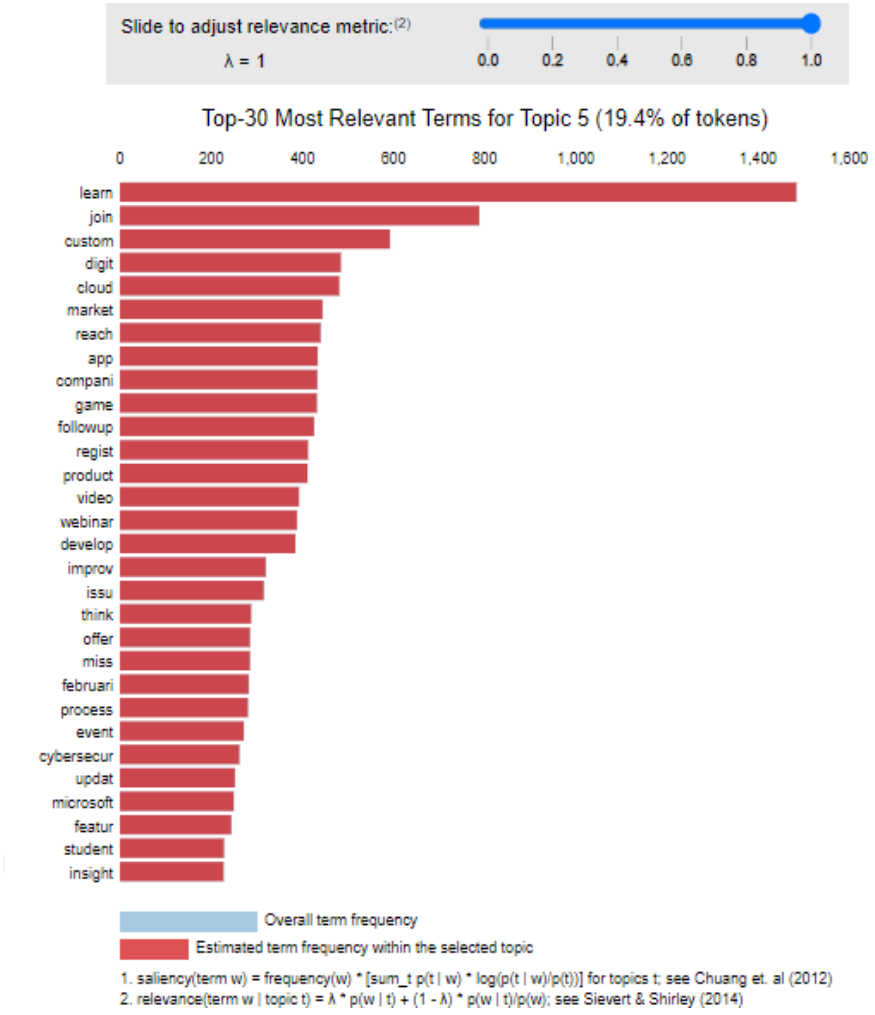


Figure 4.6 Most relevant terms for topic five

According to the results of the LDA topic modelling, the themes and concepts related to the five innovation factors in this study and the second and third research questions present all the probable topics in firms' tweets (Table 4.7). Our findings demonstrate that the first conceivable topic for this data set is creativity; the next likely topics for tweets in this data set are R&D and collaboration themes and keywords; and finally, external financing is third.

Table 4.7 Categorizing topics in dataset

| Topic | First Probable topic | Second probable topic | Third probable topic |
|---|---|---|---|
| Topic 1 | Creativity- | Collaboration -R&D | External financing |
| Topic 2 | Creativity | Collaboration - R&D | External financing |
| Topic3 | Creativity | Collaboration-R&D | External financing |
| Topic4 | External financing | Collaboration-R&D - Creativity | IP |
| Topic5 | Creativity | R&D | Collaboration |

## 4.2 Analyzing implementation of model and innovation factors as research findings

**Analyzing the implementation of model**

According to Elgesem, Steskal, and Diakopoulos (2015) and Koltsova and Shcherbak (2015), LDA topic modelling is a computational content-analysis technique that investigates the hidden thematic structure from text collection. It offers quick and efficient data-driven and computational techniques that are attractive to researchers. In fact, in this model, the inductive and quantitative measurements are combined to make it suitable for exploratory and descriptive analyses. According to Maier et al. (2018), despite increasing researchers' interest in using this model, it poses the following challenges with regards to developing good-practice guidance:(1) unstructured text data pre-processing steps, (2) selecting several topics to be generated, (3) evaluating the reliability and interpretability of the model, and finally, (4) results validation. These challenges are explained in the next part.

**Unstructured text data pre-processing steps**

According to Denny and Spirling (2017), the steps used for cleaning text can affect the input vocabulary and documents model. Therefore, to have reasonable results, it is essential to do pre-processing efficiently. According to Parra Santander (2015), the process of cleaning data is directly

related to the research question and the type of data. For example, if the research question is related to the language, the language filter should be used. Based on the idea of Ghosh and Guha (2013), in using web documents , some content such as hypertext markup language (HTML) markups needs to be removed. In using short documents ,such as tweets, the aggregation of distinct is necessary (Guo, Vargo, Pan, Ding, & Ishwar, 2016). Therefore, the standard pre-processing procedures include:(1) tokenization, (2) removing punctuation, (3) eliminating stop words, (4) steaming or lemmatizing (Lovins, 1968; Tong & Zhang, 2016).

According to this research findings, all topics have several root words and some words were in imperfect dictation due to the pre-processing steps. However, several of these words were significant because they could guide us to the innovation themes.

**Selecting the number of topics to be generated**

According to Mimno, Wallach, Talley, Leenders, and McCallum (2011), although focusing on the model parameters (e.g., α and β) that can affect the results is important , there is no standard statistical procedure to guide us for selecting correct number of topic in LDA topic modelling. It remained as one of the most complicated tasks. According to Grimmer (2010, p. 12), on the one hand , accepting too many topics can cause to have similar entities that are not meaningful , but on the other hand, as according to M. S. Evans (2014), having too few topics might result in broad entities that must then be separated based on their aspects. .

**Evaluating the reliability and interpretability of the model**

According to Mirylenka et al. (2019), two steps in the integral parts of LDA are model random initialization and sequence of multiple random processes. Hence, this is a deterministic model because of both random initialization and stochastic inference. According to Niekler and Jähnichen (2012), one of the weaknesses of the LDA is random initialization methods.

**Result validation**

According to Mirylenka et al. (2019), essential topics can be hidden in word distributions. To achieve better results, it is advisable to use a combination of different existing metrics. Using LDA allows us to extract the top n terms contributing topics. These topics are based on the model

components[27]. According to Arun, Suresh, Madhavan, and Murthy (2010), comparing the results by using different numbers of topics can help select the correct number of topics. "If the right number of topics is reached, the words belonging to every topic are expected to be semantically close to each other." (Arun et al., 2010). According to Maier et al. (2018), regard to determining an adequate number of topics ,researchers (e.g. (Biel & Gatica-Perez, 2014; Elgesem et al., 2015)) usually use various numbers of topics for the model, and models' results are compared to determine whether they are significant or not. To generate a valid topic solution, researchers consider external and internal validation criteria (Baum, 2012; M. S. Evans, 2014). Applying different metrics, such as perplexity, can help to uncover the statistical "good fit" of a topic mode, and to estimate how well a model is produced (Ghosh & Guha, 2013; Jacobi, Van Atteveldt, & Welbers, 2016). According to the McLachlan (2004), log-likelihood can be another solution.

Other scholars suggested using a nonparametric topic model, such as the Hierarchical Dirichlet Process (HDP), which does not require that the number of topics be determined, as several topics will be estimated by the data (Teh, Jordan, Beal, & Blei, 2006). Other researchers propose using parameters (which are used in LDA) and applying different values by various programming packages, such as R topic models package and Mallet software packages (David M Blei et al., 2003b; Ghosh & Guha, 2013; Hornik & Grün, 2011; McCallum, 2002).

In this research, the application of LDA topic modelling involved different steps.[28] Then, after pre-processing , we used various libraries in Python to facilitate model implementations.[29]. Firstly, using *GridSearch*[30] method in Scikit-learn package helped to values for different possible hyper parameters. As shown in Figure 4.7, different numbers of topics were tested. These are 3, 4, 5, and

---

[27] Using LDA with online variational Bayes algorithm needs to use Sklearn. Decomposition. LatentDirichletAllocation , and number of topics showed as parameters n_components.

[28] [28]Parameters are set as follows : Batch size=number of documents in each learning iter=128; Evaluate _every= compute perplexity every n iters, default= do not; Max learning iterations=10;randome state=100; N_jobe = use all available CPUS;

[29] Some libraries used for regular expressions such as (re ) for processing text (spacy and genism) for visualization (pyLDAvis and matplotlib) for manipulating (numpy and pandas)

[30]https://scikitlearn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html?highlight=gridsearch

6, with learning_decay parameters being: 0.5, 0.7, and 0.9. (Using the *learning_decay*[31] as parameters can control the rate of learning method.)

```
# Define Search Param
search_params = {'n_components': [ 6 , 5 , 4 ,  3  ], 'learning_decay': [.5, .7, .9]}

# Init the Model
lda = LatentDirichletAllocation()

# Init Grid Search Class
model = GridSearchCV(lda, param_grid=search_params)

# Do the Grid Search
model.fit(data_vectorized)
```

Figure 4.7 Implementing LDA with different number of topics

To find model performance, a number of topics were selected. To find the best model performance for all topics, two metrics were calculated (see Table 4.8): log-likelihood and perplexity. Therefore, the model that has a higher log-likelihood and lower perplexity will be considered a good model. According to the results in Table 4.8, the best log-likelihood score and perplexity are -674315.1867 and 443.83277 respectively, and they belong to topic five.

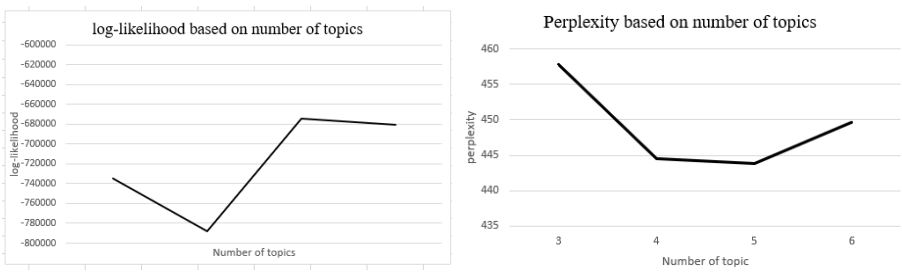


Figure 4.8 Log-likelihood and perplexity based on different number of topics

[31] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

Table 4.8 Comparing different number of topics based on their log-likelihood and perplexity

| **Number of topics** | **Log-likelihood** | **Perplexity** |
|---|---|---|
| 3 | -734621.76588245 | 457.835576 |
| 4 | -787754.92218874 | 444.534227 |
| 5 | -674315.18678263 | 443.832773 |
| 6 | -680766.328652221 | 449.664435 |

Finally, using *model.best_estimator_*[32] facilitates to find the best number of topics between these digits. Table 4.9 summarizes model information.

```
# Best Model
best_lda_model = model.best_estimator_

# Model Parameters
print("Best Model's Params: ", model.best_params_)

# Log Likelihood Score
print("Best Log Likelihood Score: ", model.best_score_)

# Perplexity
print("Model Perplexity: ", best_lda_model.perplexity(data_vectorized))
```

```
Best Model's Params:  {'learning_decay': 0.7, 'n_components': 5 }
Best Log Likelihood Score:  -674315.18678263
Model Perplexity:  443.83277365404622
```

Figure 4.9 The best model for dataset

Table 4.9 Model information

| **Model information** | **Scores** |
|---|---|
| Batch_size | 128 |
| Doc_topic_prior | None |
| Evaluate_every | -1 |
| Learning _decay | 0.7 |
| Max_doc_update_iter | 100 |

[32] https://scikit-learn.org/stable/tutorial/statistical_inference/model_selection.html

# CHAPTER 5 DISCUSSION

The following chapter presents the discussion based on our analysis of the Twitter data set and the three research questions. Section 5.1 presents social media research challenges and offers an overview of this study. The remaining sections deal with the research questions: Section 5.2 presents the first research question, and Section 5.3 covers the second and third research questions.

## 5.1 Social media research challenges

Based on Kaplan and Haenlein (2010) research, Web 2.0 helped represent the ideological and technological foundations of social media, which is characterized as a summary of user-generated content Kleinberg (2007) indicated that that there are some challenges with regards to analyzing social networks, such as (1) problems maintaining large-scale social network data for seeking and collecting information, (2) privacy issues related to using social networks. According to Bello-Orgaz, Jung, and Camacho (2016), new questions related to social media mainly focus on data processing, saving data, representation of data, finding patterns (e.g., certain behaviours), and visualization. Most of these problems are related to the unstructured nature of the data and to the quantity of data. Therefore, new big data paradigms such as Apache Hadoop with several coding libraries can be used to overcome these barriers.

**Research overview**

According to the literature, firms, like other social media users, can create an enormous amount of public information on social media, and this allows scientists to perform new research (Aral, Dellarocas, & Godes, 2013; Gayo-Avello et al., 2013; Harris, 2014). ). However, collecting, generating, and analyzing results from online microblogs and other social textual data poses some challenges, such as with the data's unstructured nature. Therefore, content analysis through natural language processing and text mining methods can be used to attempt to extract more information from these sources. Various forms of social textual data can be found, such as image descriptions, video narrations, and so forth. (Dominey & Voegtlin, 2003). This thesis aimed to investigate some information associated with the firms' Twitter accounts. Since Twitter is one of the most popular social networks, this study first looked at the firms of the Custom Computer Programming Services (NAICS code 541511) and, through web scraping, collected some features of these firms. The work

focused on innovation concepts and, using the firms' tweets, searched for some probable topics for this data set via LDA topic modelling.

**Discussion about the first research question**

The data contained in online platforms is unstructured, making its analysis a challenge. The first research question focused on firms' interest in using innovation concepts in their tweets. It is explained as follows:

RQ1 -To what extent are firms interested in innovation in their tweets?

Addressing this question required a clear understanding of innovation concepts. For the first step, we used different resources with various keywords and themes to collect our comparison list. According to Héroux-Vaillancourt et al. (2020), older Oslo Manual OECD (2005) , O. Manual (2018) and glossary of the statistical terms web page by OECD , Williams et al. (2016), and Collins Thesaurus of the English Language, five innovation factors were selected. These five factors were: R&D, IP, collaboration, external financing, and creativity. For each factor, we defined keywords and themes. Therefore, having a reference list of these words and using text mining techniques via LINQ and C# in visual studio, we answered our first research question.

**Challenges related to text mining**

According to A.-H. Tan (1999a), text mining as knowledge discovery has several pre-processing steps, and extracting information can present some patterns and knowledge regarding the subject of research. Since more than 80% of firms' information is contained in text documents[33], text mining can be used in the following areas: (1) analyzing customer profiles, (2) analyzing patents, (3) distributing information, and finally (4) resource planning. Concerning different mining productions and applications, text mining can be divided into two groups. The first group, which is document-based, relates to the visualization of the text documents; the second group is concept-based, in connection with the concept analysis functions and information or categorization. Despite its great potential, some technical issues arise with this technique, such as the need to briefly describe the documents and concepts to obtain satisfactory results. This discovery needs some

[33] https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/

semantic analysis methods, which are computationally expensive. Another problem relates to multilingual text, which requires a different syntactic structure. Collecting data, pre-processing, and using proper mining operations requires trained knowledge specialists. Some challenges remain with regards to using text data, such as pattern recognition and visualization and prediction, but improving the accuracy of these predictions is a question for scholars (Anjaria & Guddeti, 2014). Moreover, having a broad perspective about other innovation research and using other text mining techniques and applications [34] in different platforms such as programming language (e.g., python) can cause different reference lists with mixed results, as seen in this study.

## 5.2 Discussion about the second and third research question

The second and third research questions of this study are as follow:

RQ2 - What are the most frequent words used in the tweets of firms in that sector?

RQ3 - To what extent does LDA effectively identify topics related to innovation in the firm's tweets?

To answer these questions, we used NLP and LDA topic modelling to identify several common words and five critical topics used in Twitter. Furthermore, by focusing on five specific innovation factors, we were able to estimate whether these topics were related to innovation factors or not. Since our data is unlabeled text, topic modelling is a suitable approach for implementation as an unsupervised algorithm. This is an actual process of applying statistical models to extract hidden topics in the data. In fact, these findings shed light on the existence of the five innovation factors in firms' tweets. The reference tables of these five innovation factors facilitated our analysis and helped us find the innovation topics hidden in firms' tweets.

**Challenges related to applying of LDA topic modeling**

According to Tong and Zhang (2016), topic models in ML and NLP help us generate models. These models are probabilistic frameworks that are used mainly for organizing, understanding, searching,

[34] Several text mining techniques can be retrieval, extraction, summarization, categorization and clustering. Source : https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/

and summarizing large unstructured texts. The topic model discovers hidden themes throughout the collection of documents. LDA, as an NLP and statistical ML approach, is the most popular probabilistic text modelling. Applying LDA has several advantages and disadvantages. It is a probabilistic generative model that uses algorithms to train models directly, and it is a simple model to use (Girolami & Kabán, 2003). Secondly, it is suitable for large-scale text sets because the size of the parameter space of LDA is fixed; there is and no need to do anything else for the size of the text set (Masada, Kiyasu, & Miyahara, 2008). Third, hierarchical models are stable, and LDA is a hierarchical model; therefore, it is stable (Girolami & Kabán, 2003). The disadvantages of using LDA are related to dealing with mass data and processing time, which can be long (David Meir Blei, 2004). ). Hence, the implementation of large-scale text with this model meets the requirements of computation time and memory capacity (Porteous et al., 2008). Another disadvantage of this model is related to finding topic rules; therefore, using parallel LDA algorithms can help. These parallel algorithms can solve the problem of calculation and storage. Besides, many practical applications such as text classification, information retrieval, and text summarization are being enabled to address parallel topic models (David Meir Blei, 2004; Newman, Asuncion, Smyth, & Welling, 2009; Porteous et al., 2008).

**Innovation factors as research findings**

Concerning the research objectives and analysis of five innovation factors, and comparing these words with LDA topic modelling results, one of the most challenging steps in this project is extracting Twitter data topics. LDA topic modelling used a word frequency algorithm and identified probable topics for the dataset. Some words and concepts were changed in the pre-processing steps. However, looking at the roots of these words helped us determine if they were related to the innovation concepts or not.

Research findings show that keywords and themes related to creativity are most used in tweets. Based on the results, collaboration and R&D themes are the second, and external financing is the third concept used in firms' communications. Tweets use topics such as IP and their relevant words less. According to the results, extracting words showed that some words had less connection to innovation concepts, and these words refer to other topics such as time and market (e.g., day, mobile). Table 5.1, as a summary, presents innovation words according to each topic. These findings can guide the researcher to predict firms' future Twitter activities.

Table 5.1 Innovation words and themes in data set

| **Innovation keywords in topics** | | | | |
|---|---|---|---|---|
| **Topic 1** | **Topic2** | **Topic3** | **Topic4** | **Topic5** |
| Look-make New-industri-change expert-need-team-way-manag-connection-support-tip-futur-design-employ-challeng | Technology-creat-program-share-service-job-plan-open-innov | Use-know-read-work-leader-partner-organ-opportun-transform- | Order-check-experi-leader-tech- trade-right-network-power-assist-readi-secur-busi- train- | Market-app-product-develop-improv-process-join-compani-learn-think-event-featur-insight |

# CHAPTER 6 CONCLUSION AND RECOMMENDATIONS

This chapter presents the conclusion of this study, and it is divided into two sections. Section 6.1 explains some research limitations, while Section 6.2 discusses future studies.

## 6.1 Research limitations

These days, social networks are among the most interesting data mining resources for research. The sharing of opinions through microblogging services like Twitter allows researchers to access and analyze this content. This work aimed to pay attention to firms' Twitter accounts to discover innovation concepts in their tweets. Since innovation concepts play the primary role in firms and industry, this study's first question was, "to what extent are firms interested in innovation in their tweets?" Web-scraping and data-mining techniques were used to collect data. Therefore, we collected 510,979 tweets by 405 Custom Computer Programming Services sector (NAICS 541511)[35] firms to build our data set.

To achieve the research objectives, by focusing on different resources such as Héroux-Vaillancourt et al. (2020) and Williams et al. (2016) and other resources, we selected five factors (R&D, IP, collaboration, external financing, creativity) along with relevant keywords and themes. Each factor was collected as a reference table. These five reference tables helped us to analyze the tweets. Using text-mining techniques via LINQ and C# in visual studio helped us to calculate how many times firms were interested in using those factors in their tweets.

To address the second and the third research questions, "what are the most frequent words used in the tweets of firms in that sector?" and "to what extent does LDA effectively identify topics related to innovation in the firm's tweets?", NLP and LDA topic modelling helped us to visualize several relevant terms and topics used in the tweets. Furthermore, by comparing five innovation factors with LDA results, probable innovation topics have been recognized.

---

[35] Regard to appendix A, this sector is further subdivided into two subsectors: 1) software publisher (511210) and 2) computer system design services (514512).

These results showed that firms were interested in using these five factors in their tweets. For instance, firms were more interested in using creativity keywords as one of the innovation factors in their tweets, and R&D and collaboration were the second most common topics in firms' online communications. One limitation of this research was related to finding and selecting research focused on innovation concepts in social networks that we could compare with our findings. Most of the research for evaluating innovation concepts focused on webpages (Beaudry et al., 2016; Gök et al., 2015; Héroux-Vaillancourt et al., 2020)

According to Che, Safran, and Peng (2013) and Uys, Du Preez, and Uys (2008), the use of text mining and LDA topics modelling research have some limitations, as follows:

**Tools and platforms for data processing**

First of all, focusing on big data challenges, exploring the different types of hidden knowledge would require better algorithms or better tools. Using visual studio and C# (as a programming language) presented some barriers to handling big data analysis, especially the speed of analyzing unstructured data. Moreover, due to using LDA topic modelling, time of processing and analyzing is a big issue in this algorithm. Using different algorithms such as LSA can solve this problem. Furthermore, applying ML algorithms requires computer memory and capacity.

**Accuracy and trust of mining**

Since text mining and natural language processing require pre-processing steps for the unstructured text, the use of different algorithms can affect the results. Based on scholars' studies, probability in topic modelling by LDA refers to two types: (1) certain specific documents will produce a particular topic and type, (2) exact words from vocabulary collection will be produced by specific topics (Negara et al., 2019). Other models with different implementations can yield different results. Inability to calculate the correlation between topics could be another limitation of LDA.

**Multilingual text**

Using different languages with different characters, slangs, informal words, and abbreviations can affect the pre-processing steps, impacting results.

## 6.2 Future studies

As we know, topic models reveal users' interest in the tweets (He, Jia, Han, & Ding, 2014). Different subjects can be drowned via Twitter data. For example, topics can be used for labelling the Twitter data. Then, using a labelled dataset, different algorithms can be run for better performance. One of Twitter mining's common usages is emotion analysis, which has some benefits for the companies and businesses to know the needs of their customers, products, and services (Halibas, Shaffi, & Mohamed, 2018). Deep learning algorithms can classify tasks in better results. One of the challenges associated with the tweets regards the use of voluminous sets of data that cannot be handled quickly, such as using emoji. More than 40% of the tweets' writing language is informal, which makes it challenging to discover topics (Wolny, 2016).

Each tweet can include different variables, such as location, permalink, username text, date, retweets, favourites, and mentions. According to the use of topic modelling techniques, we focused on the texts. Future studies could use a combination of these variables, such as innovation topics mentioned or topics used by hashtags in tweets. Furthermore, by improving LDA topic modelling, the correlations between topics could be calculated. Focusing on the pictures based on the topics is another potential area for future research that would require image processing techniques.

# REFERENCES


Allen, T. J. (1984). Managing the flow of technology: Technology transfer and the dissemination of technological information within the R&D organization. *MIT Press Books, 1*.

Ambrose, B. (1990). An analysis of the factors affecting light industrial property valuation. *Journal of Real Estate Research, 5*(3), 355-370.

Anjaria, M., & Guddeti, R. M. R. (2014). *Influence factor based opinion mining of Twitter data using supervised learning.* Paper presented at the 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS).

Aral, S., Dellarocas, C., & Godes, D. (2013). Introduction to the special issue—social media and business transformation: a framework for research. *Information Systems Research, 24*(1), 3-13.

Arrow, K. J. (1972). Economic welfare and the allocation of resources for invention. In *Readings in industrial economics* (pp. 219-236): Springer.

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). *On finding the natural number of topics with latent dirichlet allocation: Some observations.* Paper presented at the Pacific-Asia conference on knowledge discovery and data mining.

Asghari, M., Sierra-Sosa, D., & Elmaghraby, A. (2018). *Trends on health in social media: Analysis using twitter topic modeling.* Paper presented at the 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT).

Aspasia, V., & Ourania, N. (2015). Greek food manufacturing firms' social media efforts: evidence from Facebook. *Procedia-social and behavioral sciences, 175*, 308-313.

Bae, Y., & Lee, H. (2012). Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American Society for Information Science and Technology, 63*(12), 2521-2535.

Bakshi, R. K., Kaur, N., Kaur, R., & Kaur, G. (2016). *Opinion mining and sentiment analysis.* Paper presented at the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom).

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). *Everyone's an influencer: quantifying influence on twitter.* Paper presented at the Proceedings of the fourth ACM international conference on Web search and data mining.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of management, 17*(1), 99-120.

Bates, K. A., & Flynn, E. J. (1995). *Innovation History and Competitive Advantage: A Resource-Based View Analysis of Manufacturing Technology Innovations.* Paper presented at the Academy of Management Proceedings.

Baum, D. (2012). Recognising speakers from the topics they talk about. *Speech Communication, 54*(10), 1132-1142.

Baumol, W. J. (1986). Productivity growth, convergence, and welfare: what the long-run data show. *The American economic review,* 1072-1085.

Baumol, W. J. (2002). *The free-market innovation machine: Analyzing the growth miracle of capitalism*: Princeton university press.

Beal, V. (2010). The difference between the internet and world wide web. *Webopedia, June, 24*.

Beaudry, C., Héroux-Vaillancourt, M., & Rietsch, C. (2016). *Validation of a web mining technique to measure innovation in high technology Canadian industries.* Paper presented at the CARMA 2016–1st International Conference on Advanced Research Methods and Analytics.

Bedwell, W. L., Wildman, J. L., DiazGranados, D., Salazar, M., Kramer, W. S., & Salas, E. (2012). Collaboration at work: An integrative multilevel conceptualization. *Human Resource Management Review, 22*(2), 128-145.

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45-59.

Berry, M. W. Survey of Text mining: Clustering, Classification, and Retrieval. 2004. *New York, USA: Springer Verlag. doi, 10*, 978-971.

Berry, M. W., & Kogan, J. (2010). *Text mining: applications and theory*: John Wiley & Sons.

Bertschek, I., & Kesler, R. (2017). Let the user speak: is feedback on Facebook a source of firms' innovation? *ZEW-Centre for European Economic Research Discussion Paper*(17-015).

Biel, J.-I., & Gatica-Perez, D. (2014). Mining crowdsourced first impressions in online social video. *IEEE Transactions on Multimedia, 16*(7), 2062-2074.

Bishop, C. M. (2006). *Pattern recognition and machine learning*: springer.

Blanco, G., Pérez-López, R., Fdez-Riverola, F., & Lourenço, A. M. G. (2020). Understanding the social evolution of the Java community in Stack Overflow: A 10-year study of developer interactions. *Future Generation Computer Systems, 105*, 446-454.

Blei, D. M. (2004). *Probabilistic models of text and images*: Citeseer.

Blei, D. M., & Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian analysis, 1*(1), 121-143.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003a). Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993-1022.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003b). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Bolstad, T. M., & Høili, P. F. (2019). *Influencer marketing: instagram adverts by influencers and firms: comparative effects on purchase intention, brand attitude, and word-of-mouth.*

Borisova, G., & Brown, J. R. (2013). R&D sensitivity to asset sale proceeds: New evidence on financing constraints and intangible investment. *Journal of Banking & Finance, 37*(1), 159-173.

Bortree, D. S., & Seltzer, T. (2009). Dialogic strategies and outcomes: An analysis of environmental advocacy groups' Facebook profiles. *Public relations review, 35*(3), 317-319.

Boyd-Graber, J. L., Hu, Y., & Mimno, D. (2017). *Applications of topic models* (Vol. 11): now Publishers Incorporated.

Brants, T. (2005). Test data likelihood for PLSA models. *Information Retrieval, 8*(2), 181-196.

Brown, J. R., Fazzari, S. M., & Petersen, B. C. (2009). Financing innovation and growth: Cash flow, external equity, and the 1990s R&D boom. *The Journal of Finance, 64*(1), 151-185.

Buntine, W. (2009). *Estimating likelihoods for topic models.* Paper presented at the Asian Conference on Machine Learning.

Burns, T., & Stalker, G. M. (1961). The management of innovation. London. *Tavistock Publishing. Cited in Hurley, RF and Hult, GTM (1998). Innovation, Market Orientation, and Organisational Learning: An Integration and Empirical Examination. Journal of Marketing, 62*, 42-54.

Campbell, J. C., Hindle, A., & Stroulia, E. (2015). Latent Dirichlet allocation: extracting topics from software engineering data. In *The art and science of analyzing software data* (pp. 139-159): Elsevier.

Campbell, W. M., Dagli, C. K., & Weinstein, C. J. (2013). Social network analysis with content and graphs. *Lincoln Laboratory Journal, 20*(1), 61-81.

Carbonell, P., Mayer, M. Á., & Bravo, À. (2015). *Exploring brand-name drug mentions on Twitter for pharmacovigilance.* Paper presented at the MIE.

Che, D., Safran, M., & Peng, Z. (2013). *From big data to big data mining: challenges, issues, and opportunities.* Paper presented at the International conference on database systems for advanced applications.

Chen, X., Cheng, Q., & Lo, A. K. (2013). Accounting restatements and external financing choices. *Contemporary Accounting Research, 30*(2), 750-779.

Chesbrough, H. W. (2003). *Open innovation: The new imperative for creating and profiting from technology*: Harvard Business Press.

Chuang, J., Manning, C. D., & Heer, J. (2012). *Termite: Visualization techniques for assessing textual topic models.* Paper presented at the Proceedings of the international working conference on advanced visual interfaces.

Chung, S., Animesh, A., Han, K., & Pinsonneault, A. (2014). Firm's social media efforts, consumer behavior, and firm performance.

Clark, G. E. (2009). Environmental twitter. *Environment: Science and Policy for Sustainable Development, 51*(5), 5-7.

Constantinides, E., & Fountain, S. J. (2008). Web 2.0: Conceptual foundations and marketing issues. *Journal of direct, data and digital marketing practice, 9*(3), 231-244.

Cui, A., Zhang, M., Liu, Y., Ma, S., & Zhang, K. (2012). *Discover breaking events with popular hashtags in twitter.* Paper presented at the Proceedings of the 21st ACM international conference on Information and knowledge management.

Culnan, M. J., McHugh, P. J., & Zubillaga, J. I. (2010). How large US companies can use Twitter and other social media to gain business value. *MIS Quarterly Executive, 9*(4).

Cuzzocrea, A., Loia, V., & Tommasetti, A. (2017). *Big-data-driven innovation for enterprises: innovative big value paradigms for next-generation digital ecosystems.* Paper presented at the Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics.

De Veirman, M., Cauberghe, V., & Hudders, L. (2017). Marketing through Instagram influencers: the impact of number of followers and product divergence on brand attitude. *International Journal of Advertising, 36*(5), 798-828.

Denny, M., & Spirling, A. (2017). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *When It Misleads, and What to Do about It (September 27, 2017).*

Dominey, P. F., & Voegtlin, T. (2003). *Learning word meaning and grammatical constructions from narrated video events.* Paper presented at the Proceedings of the HLT-NAACL 2003 workshop on Learning word meaning from non-linguistic data.

Dosi, G., & Nelson, R. R. (1994). An introduction to evolutionary theories in economics. *Journal of evolutionary economics, 4*(3), 153-172.

Drake, G. (2003). 'This place gives me space': Place and Creativity in the Creative Industries. *Geoforum, 34*(4), 511-524.

Driscoll, K., & Walker, S. (2014). Big data, big questions| working within a black box: Transparency in the collection and production of big twitter data. *International Journal of Communication, 8*, 20.

Elgesem, D., Steskal, L., & Diakopoulos, N. (2015). Structure and content of the discourse on climate change in the blogosphere: The big picture. *Environmental Communication, 9*(2), 169-188.

Etzkowitz, H., & Leydesdorff, L. (2000). The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. *Research Policy, 29*(2), 109-123.

Evans, D. (2010). *Social media marketing: An hour a day*: John Wiley & Sons.

Evans, M. S. (2014). A computational approach to qualitative analysis in large textual datasets. *PloS one, 9*(2), e87908.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of medical Internet research, 13*(4), e123.

Fawcett, S. E., Magnan, G. M., & McCarter, M. W. (2008). Benefits, barriers, and bridges to effective supply chain management. *Supply Chain Management: An International Journal, 13*(1), 35-48.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*: Cambridge university press.

Francis, L. A. (2006). *Taming Text: An Introduction to Text Mining.* Paper presented at the Casualty Actuarial Society Forum.

Franke, J., Nakhaeizadeh, G., & Renz, I. (2003). *Text Mining*: Physica-Verlag.

Frascati, M. (2015). Proposed Standard Practice for Surveys on Research and Experimental Development, The Measurement of Scientific, Technological and Innovation Activities. In (pp. 44-45): Paris: OECD Publishing.

Freeman, C., & Soete, L. (1997). *The economics of industrial innovation*: Psychology Press.

Gayo-Avello, D., Metaxas, P. T., Mustafaraj, E., Strohmaier, M., Schoen, H., Gloor, P., . . . Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*.

Geum, Y., Lee, S., Yoon, B., & Park, Y. (2013). Identifying and evaluating strategic partners for collaborative R&D: Index-based approach using patents and publications. *Technovation, 33*(6-7), 211-224.

Ghosh, D., & Guha, R. (2013). What are we 'tweeting'about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and geographic information science, 40*(2), 90-102.

Girolami, M., & Kabán, A. (2003). *On an equivalence between PLSI and LDA.* Paper presented at the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.

Goh, K.-Y., Heng, C.-S., & Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. *Information Systems Research, 24*(1), 88-107.

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics, 102*(1), 653-671.

Granstrand, O. (1999). The economics and management of intellectual property. *Books*.

Greenhalgh, C., & Longland, M. (2001). Intellectual property in UK firms: creating intangible assets and distributing the benefits via wages and jobs. *Oxford Bulletin of Economics and Statistics, 63*, 671-671.

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis, 18*(1), 1-35.

Gulati, R., Nohria, N., & Zaheer, A. (2000). Strategic networks. *Strategic management journal, 21*(3), 203-215.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly, 93*(2), 332-359.

Halibas, A. S., Shaffi, A. S., & Mohamed, M. A. K. V. (2018). *Application of text classification and clustering of Twitter data for business analytics.* Paper presented at the 2018 Majan International Conference (MIC).

Hall, B. H. (2002). The financing of research and development. *Oxford review of economic policy, 18*(1), 35-51.

Harris, D. (2014). 3 lessons in big data from the Ford Motor Company. In.

He, L., Jia, Y., Han, W., & Ding, Z. (2014). Mining user interest in microblogs with a user-topic model. *China Communications, 11*(8), 131-144.

Héroux-Vaillancourt, M., Beaudry, C., & Rietsch, C. (2020). Using web content analysis to create innovation indicators–What do we really measure? *Quantitative Science Studies*, 1-37.

Heunks, F. J. (1998). Innovation, creativity and success. *Small Business Economics, 10*(3), 263-272.

Hofmann, T. (1999). Probabilistic latent semantic analysis. InProceedings of the Fifteenth conference on Uncertainty in artificial intelligence (pp. 289-296). In: Morgan Kaufmann Publishers Inc.

Hofmann, T. (1999). *Probabilistic latent semantic indexing.* Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

Hong, L., & Davison, B. D. (2010). *Empirical study of topic modeling in twitter.* Paper presented at the Proceedings of the first workshop on social media analytics.

Hornik, K., & Grün, B. (2011). topicmodels: An R package for fitting topic models. *Journal of statistical software, 40*(13), 1-30.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). *A brief survey of text mining.* Paper presented at the Ldv Forum.

Howells, J. (2008). New directions in R&D: current and prospective challenges. *R&d Management, 38*(3), 241-252.

Hyun Kim, J. (2012). A hyperlink and semantic network analysis of the triple helix (University-Government-Industry): The interorganizational communication structure of nanotechnology. *Journal of Computer-Mediated Communication, 17*(2), 152-170.

Iammarino, S. (2005). An evolutionary integrated view of regional systems of innovation: concepts, measures and historical perspectives. *European planning studies, 13*(4), 497-519.

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism, 4*(1), 89-106.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). *Why we twitter: understanding microblogging usage and communities.* Paper presented at the Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis.

Jin, X., Wang, J., Chu, T., & Xia, J. (2018). Knowledge source strategy and enterprise innovation performance: dynamic analysis based on machine learning. *Technology Analysis & Strategic Management, 30*(1), 71-83.

Jónsson, E., & Stolee, J. (2015). An evaluation of topic modelling techniques for twitter. In: Toronto: University of Toronto.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons, 53*(1), 59-68.

Karanikas, H., & Theodoulidis, B. (2002). Knowledge discovery in text and text mining software. *Centre for Research in Information Management, Department of Computation.*

Kassema, J. J. (2019). Research and Development: Key Factor For IT Innovation and Creativity. *Available at SSRN 3483600.*

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology, 1*(1), 4-20.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business horizons, 54*(3), 241-251.

Kleinberg, J. M. (2007). *Challenges in mining social network data: processes, privacy, and paradoxes.* Paper presented at the Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining.

Koltsova, O., & Shcherbak, A. (2015). 'LiveJournal Libra!': The political blogosphere and voting preferences in Russia in 2011–2012. *New Media & Society, 17*(10), 1715-1732.

Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert systems with applications, 40*(10), 4065-4074.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics, 22*(1), 79-86.

Kywe, S. M., Hoang, T.-A., Lim, E.-P., & Zhu, F. (2012). *On recommending hashtags in twitter networks.* Paper presented at the International conference on social informatics.

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems, 58*, 85-96.

Leenders, R. T. A., Van Engelen, J. M., & Kratzer, J. (2003). Virtuality, communication, and new product team creativity: a social network perspective. *Journal of Engineering and technology management, 20*(1-2), 69-92.

Leonardi, P. M., Huysman, M., & Steinfield, C. (2013). Enterprise social media: Definition, history, and prospects for the study of social technologies in organizations. *Journal of Computer-Mediated Communication, 19*(1), 1-19.

Libaers, D., Hicks, D., & Porter, A. L. (2016). A taxonomy of small firm technology commercialization. *Industrial and Corporate Change, 25*(3), 371-405.

Liddy, E. D. (2001a). Natural language processing. 2.

Liddy, E. D. (2001b). Natural language processing.

Lieberman, M. (2014). *Visualizing big data: Social network analysis.* Paper presented at the Digital research conference.

Liu, B. F., Fraustino, J. D., & Jin, Y. (2016). Social media use during disasters: How information form and source influence intended behavioral responses. *Communication Research, 43*(5), 626-646.

Lohr, S. (2013). Big data, trying to build better workers. *The New York Times, 21*.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics, 11*(1-2), 22-31.

Lu, Y., Mei, Q., & Zhai, C. (2011). Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval, 14*(2), 178-203.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., . . . Häussler, T. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures, 12*(2-3), 93-118.

Majumdar, A., & Bose, I. (2019). Do tweets create value? A multi-period analysis of Twitter use and content of tweets for manufacturing firms. *International Journal of Production Economics, 216*, 1-11.

Manual, O. (2018). Guidelines for collecting, reporting and using data on innovation, The measurement of scientific, technological and innovation activities. *October, 22*, 255p.

Manual, O. F. (2013). Retrieved from https://stats.oecd.org/glossary

Masada, T., Kiyasu, S., & Miyahara, S. (2008). *Comparing LDA with pLSI as a dimensionality reduction method in document clustering.* Paper presented at the International Conference on Large-Scale Knowledge Resources.

Mathiak, B., & Eckstein, S. (2004). *Five steps to text mining in biomedical literature.* Paper presented at the Proceedings of the second European workshop on data mining and text mining in bioinformatics.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit (2002). In.

McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition* (Vol. 544): John Wiley & Sons.

Michalak, T. P., Rahwan, T., & Wooldridge, M. J. (2017). *Strategic Social Network Analysis.* Paper presented at the AAAI.

Miles, I. (2007). Research and development (R&D) beyond manufacturing: the strange case of services R&D. *R&d Management, 37*(3), 249-268.

Miller, K. D., & Bromiley, P. (1990). Strategic risk and corporate performance: An analysis of alternative risk measures. *Academy of management journal, 33*(4), 756-779.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). *Optimizing semantic coherence in topic models.* Paper presented at the Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing.

Minh, D.-V. (2013). Sentiment and Influence Analysis of Twitter Tweets. In: Google Patents.

Minka, T. P. (2013). Expectation propagation for approximate Bayesian inference. *arXiv preprint arXiv:1301.2294*.

Mirylenka, K., Scotton, P., Miksovic, C., & Dillon, J. (2019). *Hidden Layer Models for Company Representations and Product Recommendations.* Paper presented at the EDBT.

Mulkay, B., Hall, B. H., & Mairesse, J. (2001). Firm level investment and R&D in France and the United States: A comparison. In *Investing today for the world of tomorrow* (pp. 229-273): Springer.

Nallapati, R., & Cohen, W. W. (2008). *Link-PLSA-LDA: A New Unsupervised Model for Topics and Influence of Blogs.* Paper presented at the icwsm.

Negara, E. S., Triadi, D., & Andryani, R. (2019). *Topic Modelling Twitter Data with Latent Dirichlet Allocation Method.* Paper presented at the 2019 International Conference on Electrical Engineering and Computer Science (ICECOS).

Newman, D., Asuncion, A., Smyth, P., & Welling, M. (2009). Distributed algorithms for topic models. *Journal of machine Learning research, 10*(8).

Niekler, A., & Jähnichen, P. (2012). *Matching results of latent dirichlet allocation for text.* Paper presented at the Proceedings of ICCM.

Nofsinger, J. R., & Wang, W. (2011). Determinants of start-up firm external financing worldwide. *Journal of Banking & Finance, 35*(9), 2282-2294.

O'Reilly, T., & Battelle, J. (2009). *Web squared: Web 2.0 five years on*: " O'Reilly Media, Inc.".

OECD, E. (2005). Guidelines for Collecting and Interpreting Innovation Data-Oslo Manual. *Organization for Economic Co-operation and Development, European Commission Eurostat*, 9-25.

OECD/Eurostat. (2018). Oslo Manual: Guidelines for Collecting, Reporting and Using Data on Innovation.

Onay, C., & Öztürk, E. (2018). A review of credit scoring research in the age of Big Data. *Journal of Financial Regulation and Compliance*, 382.

Pak, A., & Paroubek, P. (2010). *Twitter as a corpus for sentiment analysis and opinion mining.* Paper presented at the LREc.

Parra Santander, D. A. (2015). Twitter in academic events: A study of temporal usage, communication, sentimental and topical patterns in 16 Computer Science conferences.

Pavitt, K. (1990). What we know about the strategic management of technology. *California management review, 32*(3), 17-26.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

Perry, D. C., Taylor, M., & Doerfel, M. L. (2003). Internet-based communication in crisis management. *Management communication quarterly, 17*(2), 206-232.

Pillai, A. S., & Rao, K. S. (1996). Performance monitoring in R&D projects. *R&d Management, 26*(1), 57-65.

Pisano, G. P., & Teece, D. J. (2007). How to capture value from innovation: Shaping intellectual property and industry architecture. *California management review, 50*(1), 278-296.

Plotnikova, V. (2018). *Towards a data mining methodology for the banking domain.* Paper presented at the Proceedings of the Doctoral Consortium Papers Presented at the 30th International Conference on Advanced Information Systems Engineering, CAiSE.

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). *Fast collapsed gibbs sampling for latent dirichlet allocation.* Paper presented at the Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.

Potts, J., Cunningham, S., Hartley, J., & Ormerod, P. (2008). Social network markets: a new definition of the creative industries. *Journal of cultural economics, 32*(3), 167-185.

Ramage, D., Dumais, S. T., & Liebling, D. J. (2010). Characterizing microblogs with topic models. *icwsm, 10*(1), 16.

Ratner, A., & Ré, C. (2018). Knowledge Base Construction in the Machine-learning Era. *Queue, 16*(3), 79-90.

Reitzig, M. (2004). Strategic management of intellectual property. *MIT Sloan Management Review, 45*(3), 35.

Romer, P. (2002). When should we use intellectual property rights? *American Economic Review, 92*(2), 213-216.

Rosenberg, N. (1963). Technological change in the machine tool industry, 1840-1910. *Journal of economic history*, 414-443.

Russom, P. (2011). Big data analytics. *TDWI best practices report, fourth quarter, 19*(4), 1-34.

Safko, L. (2010). *The social media bible: tactics, tools, and strategies for business success*: John Wiley & Sons.

Sag, M. (2019). The new legal landscape for text mining and machine learning. *Journal of the Copyright Society of the USA, 66*.

Samarawickrama, S., Karunasekera, S., & Harwood, A. (2015). *Finding high-level topics and tweet labeling using topic models.* Paper presented at the 2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS).

Savignac, F. (2008). Impact of financial constraints on innovation: What can be learned from a direct measure? *Econ. Innov. New Techn., 17*(6), 553-569.

Sawhney, M., Verona, G., & Prandelli, E. (2005). Collaborating to create: The Internet as a platform for customer engagement in product innovation. *Journal of interactive marketing, 19*(4), 4-17.

Scherer, F. M. (2011). *New perspectives on economic growth and technological innovation*: Brookings Institution Press.

Shao, G. (2009). Understanding the appeal of user-generated media: a uses and gratification perspective. *Internet Research*.

Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations. *PloS one, 7*(11), e47523.

Sievert, C., & Shirley, K. (2014). *LDAvis: A method for visualizing and interpreting topics.* Paper presented at the Proceedings of the workshop on interactive language learning, visualization, and interfaces.

Smaiti, M., & Hanoune, M. (2015). Big Data: Features, Architecture, Research and Applications.

Solow, R. M. (1956). A contribution to the theory of economic growth. *The quarterly journal of economics, 70*(1), 65-94.

Solow, R. M. (1957). Technical change and the aggregate production function. *The review of Economics and Statistics*, 312-320.

Song, M. (2008). *Handbook of research on text and web mining technologies*: IGI global.

Srivastava, A. N., & Sahami, M. (2009). *Text mining: Classification, clustering, and applications*: CRC Press.

Stieglitz, S., & Dang-Xuan, L. (2012). *Political communication and influence through microblogging--An empirical analysis of sentiment in Twitter messages and retweet behavior.* Paper presented at the 2012 45th Hawaii International Conference on System Sciences.

Tajvidi, R., & Karami, A. (2017). The effect of social media on firm performance. *Computers in Human Behavior*, 105174.

Tan, A.-H. (1999a). Text Mining: promises and challenges. *South East Asia Regional Computer Confederation, Sigapore*.

Tan, A.-H. (1999b). *Text mining: The state of the art and the challenges.* Paper presented at the Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases.

Tan, K. H., Zhan, Y., Ji, G., Ye, F., & Chang, C. (2015). Harvesting big data to enhance supply chain innovation capabilities: An analytic infrastructure based on deduction graph. *International Journal of Production Economics, 165*, 223-233.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association, 101*(476), 1566-1581.

Thomson, A. M., Perry, J. L., & Miller, T. K. (2008). Forthcoming. Linking collaboration processes and outcomes: Foundations for advancing empirical theory. *Collaborative public*

*management: The big questions, ed. Rosemary O'Leary and Lisa Bingham. Armonk, NY: Sharpe*.

Tong, Z., & Zhang, H. (2016). *A text mining research based on LDA topic modelling.* Paper presented at the International Conference on Computer Science, Engineering and Information Technology.

Trott, P. (2008). *Innovation management and new product development*: Pearson education.

Tucker, A. B. (2004). *Computer science handbook*: CRC press.

Utterback, J. M. (1994). Mastering the Dynamics of Innovation (Boston, MA: Harvard Business School Press).

Uys, J., Du Preez, N., & Uys, E. (2008). *Leveraging unstructured information using topic modelling.* Paper presented at the PICMET'08-2008 Portland International Conference on Management of Engineering & Technology.

Vickery, G., & Wunsch-Vincent, S. (2007). *Participative web and user-created content: Web 2.0 wikis and social networking*: Organization for Economic Cooperation and Development (OECD).

Wang, F., & Vaughan, L. (2014). Firm web visibility and its business value. *Internet Research*.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). *A system for real-time twitter sentiment analysis of 2012 us presidential election cycle.* Paper presented at the Proceedings of the ACL 2012 system demonstrations.

Wasserman, S., & Galaskiewicz, J. (1994). *Advances in social network analysis: Research in the social and behavioral sciences*: Sage.

Wedenberg, K., & Sjöberg, A. (2014). Online inference of topics: Implementation of the topic model Latent Dirichlet Allocation using an online variational bayes inference algorithm to sort news articles. In.

Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C. (2014). *Twitter and society [Digital Formations, Volume 89]*: Peter Lang Publishing.

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). *Twitterrank: finding topic-sensitive influential twitterers.* Paper presented at the Proceedings of the third ACM international conference on Web search and data mining.

Whiting, A., & Williams, D. (2013). Why people use social media: a uses and gratifications approach. *Qualitative Market Research: An International Journal*.

Wiering, M., & van Otterlo, M. (2012). Reinforcement Learning: State-of-theart, vol. 12. In: Springer Science & Business Media, New York.

Wikström, P. (2014). # srynotfunny: Communicative functions of hashtags on Twitter. *SKY Journal of Linguistics, 27*, 127-152.

Williams, R., Runco, M. A., & Berlow, E. (2016). Mapping the themes, impact, and cohesion of creativity research over the last 25 years. *Creativity Research Journal, 28*(4), 385-394.

Wolny, W. (2016). *Knowledge gained from Twitter data.* Paper presented at the 2016 Federated Conference on Computer Science and Information Systems (FedCSIS).

Wright, D. K., & Hinson, M. D. (2009). An updated look at the impact of social media on public relations practice. *Public relations journal, 3*(2), 1-27.

Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism management, 31*(2), 179-188.

Yin, Z., Fabbri, D., Rosenbloom, S. T., & Malin, B. (2015). A scalable framework to detect personal health mentions on Twitter. *Journal of medical Internet research, 17*(6), e138.

Zanasi, A. (2007). *Text mining and its applications to intelligence, CRM and knowledge management* (Vol. 7): Wit Press.

# APPENDIX A NAICS CODES INFORMATION

Table A.1 Information related to the code 514511 and two sub-divisions 511210 and 514512

| codes | Description | Yes/No | Why it selected? |
|---|---|---|---|
| 541511 | 541511<br>custom computer programming services<br>511210<br>Software Publishers<br>514512<br>Computer system design services | yes | It covers: Applications software programming services, custom computer-Computer program or software development-Computer programming services-Computer software analysis and design services-Computer software programming services-Computer software support services-Programming services-Software analysis and design services-Software programming services-Web (i.e., Internet) page design services |
| 511210 | 511210<br>Software publisher<br>42<br>wholesale<br>423430<br>Computer and computer peripheral equipment and software merchant wholesalers<br>423690<br>Electronic parts in wholesalers<br>541512<br>Computer system design services<br>44-45<br>Retail trade<br>443141<br>Electronic stores<br>444130<br>Hardware stores<br>518210<br>Data processing, hosting and related services (many codes)<br>541511<br>Customs computer programming services(repeat)<br>334614<br>Software and other pre-recorded compact disc, tape, and record reproducing<br>(512 and 334613) | yes<br>yes | It covers: Applications development and publishing, except on a custom basis-applications software-computer packaged-computer software publishers packaged-computer software reproduction-games-computer software-operating systems packaged-programming language and compiler software publisher package-utility software |

Table A.2 Information about code 514512

| Code | Description | | Yes/no | Why it selected? |
|---|---|---|---|---|
| 514512 | 514512<br>Computer system design services | | yes | It covers: computer-aided design, aided engineering, aided manufacturing, system integration –consulting services-analysing and integrator services-network system integration design services |
| | 443142 electronic stores | 423430<br>Computer and computer peripheral equipment | | |

# APPENDIX B TOPIC MODELLING AND TEXT MINING

```python
import os
class configurations():
    def __init__(self):
        self.SinceDate = "2019-02-01"
        self.UntilDate = "2019-03-01"
        self.path_output = os.path.join(os.getcwd(), "Output_2020506") # pay attention: the folder should exist before running the code!
        ## Create folder if it does not exist
        if os.path.exists(self.path_output):
            print("INFO: The output path already exists!")
        else:
            print("Warning: the output path does not exists! We create it for you...")
            os.makedirs(self.path_output)
            if os.path.exists(self.path_output):
                print("... the output path now exists!")
            else:
                print("Error: some errors occurred in creating the path, please check the code! ")
```

Figure B.1 Codes for collecting companies' information


```
[{"company": "Absorb Software Inc", "tweet_text_comp": "\n;2019-02-28 12:30;0;0;\"Keep your business edge! Build a data-backed strategy that connects learner data to core business metrics. Start by learning the strategic benefits of adding business intelligence to your LMS. https://absorbl.ms/2TJhuRX #BI #elearning #L &D pic.twitter.com/uCO4TO7EjA\";;;#BI #elearning #L;\"1101172695113388033\";https://twitter.com/AbsorbLMS/status/1101172695113388033", "tweet_username": "", "tweet_date": "2019-02-28 12:30", "tweet_retweets": 0, "tweet_favorites": 0, "tweet_text": "Keep your business edge! Build a data-backed strategy that connects learner data to core business metrics. Start by learning the strategic benefits of adding business intelligence to your LMS. https:// absorbl.ms/2TJhuRX #BI #elearning #L &D pic.twitter.com/uCO4TO7EjA", "tweet_geo_location": "", "tweet_mentions": "", "tweet_hashtags": "#BI #elearning #L", "tweet_id": "1101172695113388033", "tweet_permalink": "https://twitter.com/AbsorbLMS/status/1101172695113388033"}, {"company": "Absorb Software Inc", "tweet_text_comp": "\n;2019-02-28 08:05;0;0;\"Want a winning business? Build a winning team! Uncover the formula top companies use to recruit\u2013and retain\u2013top talent. Get the eBook to make it happen in 2019. #LMS #eLearning #eBook https:// absorbl.ms/2DDPrfO pic.twitter.com/ipmujbKkjk\";;;#LMS #eLearning #eBook;\"1101105992799002624\";https://twitter.com/AbsorbLMS/status/1101105992799002624", "tweet_username": "", "tweet_date": "2019-02-28 08:05", "tweet_retweets": 0, "tweet_favorites": 0, "tweet_text": "Want a winning business? Build a winning team! Uncover the formula top companies use to recruit\u2013and retain\u2013top talent. Get the eBook to make it happen in 2019. #LMS #eLearning #eBook https:// absorbl.ms/2DDPrfO pic.twitter.com/ipmujbKkjk", "tweet_geo_location": "", "tweet_mentions": "", "tweet_hashtags": "#LMS #eLearning #eBook",
```


Figure B.2 A JSON file sample for a company

```
In [2]: import pandas as pd
        import numpy as np
        import seaborn as sns
        import matplotlib.pyplot as plt
        from sklearn.feature_extraction.text import CountVectorizer
        import nltk
        import string
        import re
        %matplotlib inline
        pd.set_option('display.max_colwidth', 100)

In [3]: def load_data():
            data = pd.read_csv("D:data2.csv")
            return data

In [4]: tweet_df = load_data()
        tweet_df.head()
```

Figure B.3 Importing libraries and data set in the Jupiter notebook

```
tweet_df = load_data()
tweet_df.head()
```

| | company | tweet_text_comp | tweet_username | tweet_date | tweet_retweets | tweet_favorites | tweet_text | tweet_geo_location | tweet_mentions | tweet_ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Absorb Software Inc | \r\n;2019-02-28 12:30;0;0;"Keep your business edge! Build a data-backed strategy that connects l... | NaN | 2019-02-28 12:30 | 0 | 0 | Keep your business edge! Build a data-backed strategy that connects learner data to core busines... | NaN | NaN | #BI # |
| 1 | Absorb Software Inc | \r\n;2019-02-28 08:05;0;0;"Want a winning business? Build a winning team! Uncover the formula to... | NaN | 2019-02-28 08:05 | 0 | 0 | Want a winning business? Build a winning team! Uncover the formula top companies use to recruit–... | NaN | NaN | # |
| 2 | Absorb Software Inc | \r\n;2019-02-27 14:04;0;0;"Learn how to propel employee productivity from the starting blocks to... | NaN | 2019-02-27 14:04 | 0 | 0 | Learn how to propel employee productivity from the starting blocks to the finish line in this in... | NaN | NaN | # #Train |
| 3 | Absorb Software Inc | \r\n;2019-02-27 11:02;0;0;"What to do when you no longer love your | NaN | 2019-02-27 11:02 | 0 | 0 | What to do when you no longer love your #LMS ? Find out in this blog: https:// | NaN | NaN | #LMS |

Figure B.4 Data description

```
print('Dataset size:',tweet_df.shape)
print('Columns are:',tweet_df.columns)
```

```
Dataset size: (25989, 12)
Columns are: Index(['company', 'tweet_text_comp', 'tweet_username', 'tweet_date',
       'tweet_retweets', 'tweet_favorites', 'tweet_text', 'tweet_geo_location',
       'tweet_mentions', 'tweet_hashtags', 'tweet_id', 'tweet_permalink'],
      dtype='object')
```

Figure B.5 Data description

```
tweet_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25989 entries, 0 to 25988
Data columns (total 12 columns):
company               25989 non-null object
tweet_text_comp       25989 non-null object
tweet_username        0 non-null float64
tweet_date            25989 non-null object
tweet_retweets        25989 non-null int64
tweet_favorites       25989 non-null int64
tweet_text            25923 non-null object
tweet_geo_location    0 non-null float64
tweet_mentions        9019 non-null object
tweet_hashtags        10161 non-null object
tweet_id              25989 non-null int64
tweet_permalink       25989 non-null object
dtypes: float64(2), int64(3), object(7)
memory usage: 2.4+ MB
```

Figure B.6 Information about the size of data set

```
df = df.replace(np.nan, '', regex=True)
def remove_punct(text):
    print(text)
    text  = "".join([char for char in text if char not in string.punctuation])
    text = re.sub('\d', '', text)
    return text

df['Tweet_punct'] = df['tweet_text'].apply(lambda x: remove_punct(x))
df.head(10)
```

```
g eLearning learningexperience pictwittercomueMIOcqf
Its a lot easier to meet your business goals when your team is on the same page Learn how successful businesses use LMS t
echnology to keep employee and company goals aligned eLearning training https absorblmsDyMrkS pictwittercomlmDczAmih
Nothing says progress like a rewarding new career Find yours and apply at https absorblmsByZin hiring eLearning Bostonjob
s YYCjobs pictwittercomfhhqEVJp
Empower your employees to reach new heights Incorporate learning in the flow of work https absorblmstPqf LMS eLearning Ab
sorbInfuse pictwittercomyMtKAfqr
What an amazing  days Thank you to everyone who stopped by and had the chance to connect with our team LTUK pictwittercom
OdCrZiqiK
Concerned about the security of your LMS data Dont worry weve got you covered Were SOC  compliant and take your trust ver
y seriously https absorblmsBrFIt soccertified compliance datasecurity soc lms pictwittercomCXpQIPhBQ
Ready to see the latest and greatest that Absorb has to offer its UK and EU customers Introducing Business Intelligence w
ith Stephen Miller from  in the Demo Zone at LTUK pictwittercomtkdIiyQjvE
Join Richard Nantel as he leads a discussion around why your sales team might need a different training strategy today in
Theatre  of the LTUK Expo Hall from  pictwittercomDXjycMGb
Were all set for Day  of LTUK Our team of experts are here all day giving handson demos and showcasing the latest Absorb
```

Figure B.7 Removing punctuation function and its result

```
def tokenization(text):
    text = re.split('\W+', text)
    return text

df['Tweet_tokenized'] = df['Tweet_punct'].apply(lambda x: tokenization(x.lower()))
df.head()
```

| | tweet_id | tweet_text | Tweet_punct | Tweet_tokenized |
|---|---|---|---|---|
| 0 | 1101172695113388033 | Keep your business edge Build a databacked strategy that connects learner data to core business ... | Keep your business edge Build a databacked strategy that connects learner data to core business ... | [keep, your, business, edge, build, a, databacked, strategy, that, connects, learner, data, to, ... |
| 1 | 1101105992799002624 | Want a winning business Build a winning team Uncover the formula top companies use to recruitand... | Want a winning business Build a winning team Uncover the formula top companies use to recruitand... | [want, a, winning, business, build, a, winning, team, uncover, the, formula, top, companies, use... |
| 2 | 1100834116273090561 | Learn how to propel employee productivity from the starting blocks to the finish line in this in... | Learn how to propel employee productivity from the starting blocks to the finish line in this in... | [learn, how, to, propel, employee, productivity, from, the, starting, blocks, to, the, finish, l... |
| 3 | 1100788364616364032 | What to do when you no longer love your LMS Find out in this blog https absorblmsTMjaz Training... | What to do when you no longer love your LMS Find out in this blog https absorblmsTMjaz Training... | [what, to, do, when, you, no, longer, love, your, lms, find, out, in, this, blog, https, absorbl... |
| 4 | 1100428091388760065 | Reporting tools bridge the gap between data and action setting you up for successand major kudos... | Reporting tools bridge the gap between data and action setting you up for successand major kudos... | [reporting, tools, bridge, the, gap, between, data, and, action, setting, you, up, for, successa... |

Figure B.8 Tokenization function and its result

```
nltk.download('stopwords')
stopword = nltk.corpus.stopwords.words('english')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Melika\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
def remove_stopwords(text):
    text = [word for word in text if word not in stopword]
    return text

df['Tweet_nonstop'] = df['Tweet_tokenized'].apply(lambda x: remove_stopwords(x))
df.head(10)
```

| | tweet_id | tweet_text | Tweet_punct | Tweet_tokenized | Tweet_nonstop |
|---|---|---|---|---|---|
| 0 | 1101172695113388033 | Keep your business edge Build a databacked strategy that connects learner data to core business ... | Keep your business edge Build a databacked strategy that connects learner data to core business ... | [keep, your, business, edge, build, a, databacked, strategy, that, connects, learner, data, to, ... | [keep, business, edge, build, databacked, strategy, connects, learner, data, core, business, met... |
| 1 | 1101105992799002624 | Want a winning business Build a winning team Uncover the formula top companies use to recruitand... | Want a winning business Build a winning team Uncover the formula top companies use to recruitand... | [want, a, winning, business, build, a, winning, team, uncover, the, formula, top, companies, use... | [want, winning, business, build, winning, team, uncover, formula, top, companies, use, recruitan... |
| 2 | 1100834116273090561 | Learn how to propel employee productivity from the starting blocks to the finish line in this in... | Learn how to propel employee productivity from the starting blocks to the finish line in this in... | [learn, how, to, propel, employee, productivity, from, the, starting, blocks, to, the, finish, l... | [learn, propel, employee, productivity, starting, blocks, finish, line, infographic, fast, track... |

Figure B.9 Using NLTK and removing stop words function and its result

```
ps = nltk.PorterStemmer()

def stemming(text):
    text = [ps.stem(word) for word in text]
    return text

df['Tweet_stemmed'] = df['Tweet_nonstop'].apply(lambda x: stemming(x))
df.head()
```

| | tweet_id | tweet_text | Tweet_punct | Tweet_tokenized | Tweet_nonstop | Tweet_stemmed |
|---|---|---|---|---|---|---|
| 0 | 1101172695113388033 | Keep your business edge Build a databacked strategy that connects learner data to core business ... | Keep your business edge Build a databacked strategy that connects learner data to core business ... | [keep, your, business, edge, build, a, databacked, strategy, that, connects, learner, data, to, ... | [keep, business, edge, build, databacked, strategy, connects, learner, data, core, business, met... | [keep, busi, edg, build, databack, strategi, connect, learner, data, core, busi, metric, start, ... |
| 1 | 1101105992799002624 | Want a winning business Build a winning team Uncover the formula top companies use to recruitand... | Want a winning business Build a winning team Uncover the formula top companies use to recruitand... | [want, a, winning, business, build, a, winning, team, uncover, the, formula, top, companies, use... | [want, winning, business, build, winning, team, uncover, formula, top, companies, use, recruitan... | [want, win, busi, build, win, team, uncov, formula, top, compani, use, recruitand, retaintop, ta... |
| 2 | 1100834116273090561 | Learn how to propel employee productivity from the starting blocks to the finish line in this in... | Learn how to propel employee productivity from the starting blocks to the finish line in this in... | [learn, how, to, propel, employee, productivity, from, the, starting, blocks, to, the, finish, l... | [learn, propel, employee, productivity, starting, blocks, finish, line, infographic, fast, track... | [learn, propel, employe, product, start, block, finish, line, infograph, fast, track, product, t... |
| 3 | 1100788364616364032 | What to do when you no longer love your LMS Find out in this blog https absorblmsTMjaz | What to do when you no longer love your LMS Find out in this blog https absorblmsTMjaz | [what, to, do, when, you, no, longer, love, your, lms, find, out, in, this, blog, https, absorbl... | [longer, love, lms, find, blog, https, absorblmstmjaz, training, learning, pictwittercomkyfrugpjul | [longer, love, lm, find, blog, http, absorblmstmjaz, train, learn, pictwittercomkyfrugpju] |

Figure B.10 Stemming the text and its result

```
nltk.download('wordnet')
wn = nltk.WordNetLemmatizer()

def lemmatizer(text):
    text = [wn.lemmatize(word) for word in text]
    return text

df['Tweet_lemmatized'] = df['Tweet_nonstop'].apply(lambda x: lemmatizer(x))
df.head()
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\Melika\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

| | tweet_id | tweet_text | Tweet_punct | Tweet_tokenized | Tweet_nonstop | Tweet_stemmed | Tweet_lemmatized |
|---|---|---|---|---|---|---|---|
| 0 | 1101172695113388033 | Keep your business edge Build a databacked strategy that connects learner data to core business ... | Keep your business edge Build a databacked strategy that connects learner data to core business ... | [keep, your, business, edge, build, a, databacked, strategy, that, connects, learner, data, to, ... | [keep, business, edge, build, databacked, strategy, connects, learner, data, core, business, met... | [keep, busi, edg, build, databack, strategi, connect, learner, data, core, busi, metric, start, ... | [keep, business, edge, build, databacked, strategy, connects, learner, data, core, business, met... |
| 1 | 1101105992799002624 | Want a winning business Build a winning team Uncover the formula top companies use to recruitand... | Want a winning business Build a winning team Uncover the formula top companies use to recruitand... | [want, a, winning, business, build, a, winning, team, uncover, the, formula, top, companies, use... | [want, winning, business, build, winning, team, uncover, formula, top, companies, use, recruitan... | [want, win, busi, build, win, team, uncov, formula, top, compani, use, recruitand, retaintop, ta... | [want, winning, business, build, winning, team, uncover, formula, top, company, use, recruitand,... |

Figure B.11 Lemmatizing the text and its result

```
def clean_text(text):
    text_lc = "".join([word.lower() for word in text if word not in string.punctuation]) # remove puntuation
    text_rc = re.sub('[0-9]+', '', text_lc)
    tokens = re.split('\W+', text_rc)
    # tokenization
    text = [ps.stem(word) for word in tokens if word not in stopword]  # remove stopwords and stemming
    return text
```

```
from sklearn.feature_extraction.text import CountVectorizer
import numpy as np
```

Figure B.12 Cleaning the text and count vectorising it