

**SCIENCE, TECHNOLOGY AND INNOVATION-RELATED TEXT DATA  
ANALYSIS WITH DEEP NEURAL NETWORKS**



**NECİP GÖZÜAÇIK**

**JUNE 2022**

**SCIENCE, TECHNOLOGY AND INNOVATION-RELATED TEXT DATA  
ANALYSIS WITH DEEP NEURAL NETWORKS**

**A PhD THESIS SUBMITTED TO THE**

**GRADUATE SCHOOL**

**OF**

**BAHÇEŞEHİR UNIVERSITY**

**BY**

**NECİP-GÖZÜAÇIK**

**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS**

**FOR**

**THE PhD DEGREE OF COMPUTER ENGINEERING**

**IN THE DEPARTMENT OF COMPUTER ENGINEERING**

**HAZİRAN 2022**



**REPUBLIC OF TURKEY  
BAHÇEŞEHİR UNIVERSITY  
GRADUATE SCHOOL**

**PhD THESIS APPROVAL FORM**


<b>Name Surname</b>	NECİP GÖZÜAÇIK
<b>Student Number</b>	1701246
<b>Program Name</b>	Computer Engineering
<b>Title of Thesis</b>	Science, Technology and Innovation-Related Text Data Analysis with Deep Neural Networks
<b>Thesis Defense Date</b>	28.06.2022

It has been approved by the Graduate School that this thesis has fulfilled the necessary conditions as a PhD thesis.

**Prof. Dr. Ahmet ÖNCÜ**  
**Director of Graduate School**

This Thesis has been read by us, it has been deemed sufficient and accepted as a PhD thesis in terms of quality and content.

<b>PhD Thesis Defense Jury</b>		
<b>Thesis Defense Jury</b>	<b>Title - Name / Surname</b>	<b>Signature</b>
Thesis Advisor	Assoc. Prof. C. Okan Şakar	
Member of Thesis Monitoring Committee	Assoc. Prof. Tevfik Aytekin	
Member of Thesis Monitoring Committee	Asst. Prof. Hamza Osman İlhan	
Member	Prof. M. Alper Tunga	
Member	Asst. Prof. Alper Özcan	



**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

Name, Surname : Necip Gözüaık

Signature :

## ABSTRACT

### SCIENCE, TECHNOLOGY AND INNOVATION-RELATED TEXT DATA ANALYSIS WITH DEEP NEURAL NETWORKS

Gözüaçık, Necip

Computer Engineering PhD Program

Supervisor: Assoc. Prof. C. Okan Şakar

June 2022, 135 pages

This thesis deals with utilizing deep neural network architectures for text data analysis of science, technology, and innovation topics. The amount of published scientific publications continues to rise fast in the Internet age. That is driven by awareness of the importance of research in the advancement of knowledge and its application through innovation. At this point, analyzing past data is important to provide a path regarding collaboration of science, technology, and innovation for the future. In this thesis, there are two main objectives. The first objective is to perform opinion mining from social media for an innovative technology product using multi-task deep neural networks with the aim of examining the effectiveness of deep learning techniques in social media mining. The second objective is to propose an end-to-end framework for technological forecasting for a selected technology domain based on the estimation of future word embedding matrix. For this purpose, the word embedding matrices throughout the years are obtained in an online learning fashion and an LSTM-based deep neural network architecture is used to model the temporal characteristics of the generated word embedding matrices and predict the future word embedding matrix accordingly. Apart from methodological contributions, this thesis offers practical contributions for the opinion mining and technological forecasting tasks regarding different use-cases and domains.

**Keywords:** Deep Learning, Natural Language Processing, Opinion Forecasting, Opinion Mining, Technology Foresighting

## ÖZ

### DERİN SİNİR AĞLARI İLE BİLİM, TEKNOLOJİ VE İNOVASYON İLE İLGİLİ METİN ANALİZİ

Gözüaçık, Necip

Computer Engineering PhD Program

Supervisor: Doç. Dr. C. Okan Şakar

Haziran 2022, 135 sayfa

Bu tez, bilim, teknoloji ve inovasyon ile ilgili metin verilerinin analizi için derin sinir ağı mimarilerinin kullanılmasıyla ilgilidir. İnternet çağında, yayınlanan bilimsel yayınların miktarı hızla artmaya devam ediyor. Bu, bilginin ilerlemesinde araştırmanın öneminin ve inovasyon yoluyla uygulanmasının farkındalığından kaynaklanmaktadır. Bu noktada, geçmiş verilerin analiz edilmesi, geleceğe yönelik bilim, teknoloji ve inovasyon alanında iş birliklerinin öngörülmesi açısından önemlidir. Bu tezde iki temel hedef vardır. İlk hedef, sosyal medya madenciliğinde derin öğrenme tekniklerinin etkinliğini incelemek amacıyla, çok görevli derin sinir ağlarını kullanarak yenilikçi bir teknoloji ürünü için sosyal medyadan fikir madenciliği yapmaktır. İkinci hedef, kelime matrisinin tahminine dayalı olarak seçilen bir teknoloji alanı için uçtan uca bir mimari önermektir. Bu amaçla, yıllar bazında kelime matrisleri çevrimiçi öğrenme şeklinde elde edilmiş ve oluşturulan kelime matrislerinin zamansal özelliklerini modellemek ve buna göre gelecekteki kelime matrisini tahmin etmek için LSTM tabanlı bir derin sinir ağı mimarisi kullanılmıştır. Metodolojik katkıların yanı sıra, bu tez, farklı kullanım senaryoları ve alanları ile ilgili fikir madenciliği ve teknolojik tahmin görevleri için pratik katkılar sunmaktadır.

**Anahtar Kelimeler:** Derin Öğrenme, Doğal Dil İşleme, Fikir Tahmini, Fikir Madenciliği, Teknoloji Öngörüsü



To my “dear” wife Şule, my mother Sevciye, my father Kadir Sami, my brother Namık Kemal  
and my sister Necla

## ACKNOWLEDGMENTS

I wish to express my deepest gratitude to my principal supervisor Assoc. Prof. C. Okan ŞAKAR for his guidance, advice, criticism, encouragements and insight throughout the research. In addition to, I wish to express my deepest gratitude to my co-supervisor Assoc. Prof. Sercan ÖZCAN. Lastly, I thank to all the members of my thesis supervising committee Assoc. Prof. Tevfik AYTEKİN, Asst. Prof. Hamza Osman İLHAN for their constructive feedbacks and guidance.

I would also like to thank Asst. Prof. Oğuzhan YAVUZ, Ayşe Belma KAYA and Caner AKSOY for their encouragement to start PhD program. Özde TİRYAKİ also helped me a lot during preparation to PhD qualification exam. I would also like to thank Bahadır ÖZDEMİR for his support and guidance during my professional career.

I would also like to thank my wife, Şule GENÇ GÖZÜAÇIK, for her great support. I would never have been able to pursue this level of school and finish this PhD without her compassion and unwavering support.



## TABLE OF CONTENTS

ETHICAL CONDUCT .....	iii
ABSTRACT.....	iv
ÖZ .....	v
DEDICATION .....	vi
ACKNOWLEDGMENTS.....	vii
TABLE OF CONTENTS .....	viii
LIST OF TABLES .....	x
LIST OF FIGURES.....	xi
LIST OF SYMBOLS/ABBREVIATIONS .....	xiii
Chapter 1: Introduction .....	1
1.1 Motivation.....	1
1.2 Research Goal .....	5
1.3 Thesis Organization .....	8
Chapter 2: Methods .....	10
2.1 Systematic Literature Review .....	10
2.1.1 Opinion Mining for Analysis of Product .....	10
2.1.2 Technological Forecasting for Technology and Innovation .....	16
2.2 Case Study .....	19
2.2.1 Opinion Mining.....	19
2.2.2 Technological Forecasting .....	20
2.3 Natural Language Processing .....	20
2.4 Text Representation Algorithms .....	22
2.4.1 Bag of Words .....	22
2.4.2 Word Embeddings.....	25
2.5 Deep Learning Algorithms.....	32
2.5.1 Artificial Neural Networks.....	32
2.5.2 Long Short-Term Memory .....	36
Chapter 3: Models .....	39
3.1 Opinion Mining for Technology Analysis with Deep Neural Networks	39
3.1.1 Dataset.....	41

3.1.2 Pre-processing .....	45
3.1.3 Feature Extraction .....	45
3.1.4 Model Creation.....	46
3.1.5 Co-occurrence Matrix-Based Visualization .....	49
3.1.6 Results .....	51
3.1.6.1 Experimental Results .....	51
3.1.6.2 Practical Results .....	53
3.2 Technology Forecasting for Technology Analysis with Deep Neural Networks .....	56
3.2.1 Dataset.....	57
3.2.2 Pre-processing .....	77
3.2.3 Feature Extraction .....	78
3.2.4 Model Creation.....	79
3.2.5 Visualization and Quantitative Analysis .....	81
3.2.6 Results .....	86
3.2.6.1 Experimental Setup .....	88
3.2.6.2 Quantitative Results .....	89
3.2.6.3 Practical Results .....	91
Chapter 4: Conclusions .....	103
4.1 Discussions.....	103
4.1.1 Opinion Mining.....	103
4.1.2 Technological Forecasting .....	104
4.2 Contributions.....	105
4.2.1 Opinion Mining.....	106
4.2.2 Technological Forecasting .....	108
4.3 Limitations and Future Works .....	110
4.3.1 Opinion Mining.....	110
4.3.2 Technological Forecasting .....	110
REFERENCES.....	111

## LIST OF TABLES

### TABLES

Table 1 Summary of the Related Works.....	12
Table 2 Summary of the Related Works.....	18
Table 3 Calculation of TF Values.....	23
Table 4 Calculation of IDF Values.....	24
Table 5 Calculation of TF-IDF Values.....	24
Table 6 Relation between Biological Neuron and Artificial Neuron .....	34
Table 7 Class Distribution of Labelled Tweets .....	45
Table 8 Exemplary Labelled Tweets .....	45
Table 9 Average Performance Results on the Test Set for the Sentiment Prediction Task.....	52
Table 10 Average Performance Results on the Test Set for the Opinion Detection Task.....	53
Table 11 Class Distribution of the Tweets Labelled with the Best Multi-Task Model .....	53
Table 12 WoS Queries.....	59
Table 13 Word2Vec Training Statistics .....	79
Table 14 Python Package List .....	87
Table 15 Optimal values of the hyperparameters for the Word2Vec and LSTM Model .....	89
Table 16 Performance Metric Comparison Using Different Prediction Horizons for 2021 .....	90
Table 17 Comparison of Clusters .....	92

## LIST OF FIGURES

### FIGURES

Figure 1 Distribution Of Computer Science Articles In WoS .....	1
Figure 2 Components Of NLP (Chowdhary, 2020).....	21
Figure 3 A PCA Projection Of Word Embedding Data.....	28
Figure 4 Nearest Neighbor Points To Label/Word “Health” .....	29
Figure 5 CBoW Model (Rong, 2014) .....	30
Figure 6 Skip-Gram Model (Rong, 2014) .....	30
Figure 7 Sample Neural Network Representation For Word2Vec .....	31
Figure 8 Actual Probabilities Table (NLP Stanford, 2002) .....	31
Figure 9 Machine Learning Tasks (Kayte, 2021) .....	33
Figure 10 Biological Neuron (Wikipedia, 2022) .....	33
Figure 11 Simple Perceptron .....	35
Figure 12 Multilayer Perceptron.....	35
Figure 13 Simple RNN Architecture .....	37
Figure 14 A Simple LSTM Architecture .....	37
Figure 15 Methods And Operations Applied To Implement Opinion Retrieval System.....	39
Figure 16 Distribution Of Tweets.....	42
Figure 17 Word Cloud For Text .....	43
Figure 18 Word Cloud For Hashtags.....	44
Figure 19 Single-Task DNN Architecture For Sentiment Target.....	47
Figure 20 Single-Task DNN Architecture For Opinion Target .....	47
Figure 21 Multi-Task DNN Architecture Using Word2Vec + GloVe Features Together.....	48
Figure 22 Co-Occurrence Matrix For Negative Group Words .....	50
Figure 23 Co-Occurrence Matrix For Positive-Neutral Group Words .....	50
Figure 24 Word Cluster Visualization Of The Tweets That Includes Opinion And Is Labelled As Negative By The Sentiment Prediction Model .....	54
Figure 25 Word Cluster Visualization Of The Tweets That Includes Opinion Labelled As Positive Or Neutral By The Sentiment Prediction Model .....	55
Figure 26 General Framework.....	57
Figure 27 Text Mining Techniques And Inter-Relationship (Talib Et Al., 2016).....	59
Figure 28 Distribution Of Publication Count .....	67
Figure 29 Number Of Articles Per Month .....	68
Figure 30 Number Of Articles Per Year.....	69

Figure 31 Number Of Documents Per Type.....	70
Figure 32 Cited Reference Count .....	71
Figure 33 Distribution Of Number Of Pages.....	72
Figure 34 Word Cloud For ‘Article Title’ .....	73
Figure 35 Word Cloud For ‘Abstract’ .....	74
Figure 36 Word Cloud For ‘Author Keywords’ .....	75
Figure 37 Word Cloud For ‘Keywords Plus’ .....	76
Figure 38 Word Cloud For ‘AT_A_AK_KP’ .....	77
Figure 39 Deep Learning Architecture .....	80
Figure 40 Distribution Of Vocabulary Set Based On N-Gram.....	81
Figure 41 Vector Representation (Tellez Et Al., 2017) .....	82
Figure 42 Screenshot Of Gephi .....	83
Figure 43 Scaled Cosine Similarity Matrix For 2018 December Actual.....	85
Figure 44 Scaled Cosine Similarity Matrix For 2021 December Actual.....	86
Figure 45 Scaled Cosine Similarity Matrix For 2021 December Predict .....	86
Figure 46 Scaled Cosine Similarity Matrix For 2024 December Predict .....	86
Figure 47 Hardware Of Supercomputer Server (UHeM, 2022) .....	87
Figure 48 Loss Versus Epoch Performance.....	91
Figure 49 2021 December Actual .....	93
Figure 50 2018 December Actual .....	94
Figure 51 2021 December Predict .....	95
Figure 52 2024 December Predict .....	98
Figure 53 2027 December Predict .....	100
Figure 54 2030 December Predict .....	102

## LIST OF SYMBOLS/ABBREVIATIONS

AI	Artificial Intelligence
AMI	Adjusted Mutual Information
ANN	Artificial Neural Networks
ARFIMA	Autoregressive Fractional Integrated Moving Average
ARI	Adjusted Rand Index
ARIMA	Autoregressive Integrated Moving Average
Bi-LSTM	Bidirectional LSTM
BoW	Bag of Words
CBow	Continuous BoW
CCA	Canonical Correlation Analysis
CNN	Convolutional Neural Network
DF	Document Frequency
DNN	Deep Neural Networks
ForSTI	Foresight for STI
HPV	Human Papillomavirus
ICT	Information and Communications Technology
IDF	Inverse Document Frequency
IoT	Internet of Things
IPO	Initial Public Offering
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing

LSTM	Long Short-Term Memory
MI	Mutual Information
MLP	Multi-Layer Perceptron
NER	Named Entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NPD	New Product Development
PCA	Principal Component Analysis
POS	Part-of-Speech
PYPI	Python Package Index
QA	Question-answering
R&D	Research and Development
RI	Rand Index
RNN	Recurrent Neural Networks
STA-LSTM	Spatio-Temporal Attention LSTM
STI	Science, Technology and Innovation
SVD	Singular Value Decomposition
SVR	Support Vector Regression
TF	Term Frequency
UGC	User-Generated Content
UIC	University-Industry Collaboration
UK	United Kingdom
US	United States
WoS	Web of Science

## Chapter 1

### Introduction

#### 1.1 Motivation

The amount of published scientific publications continues to rise fast in the Internet age. It is driven by awareness of the importance of research in the advancement of knowledge and its application through innovation (Filippov & Hofheinz, 2016). For example, Google Scholar is a research database that indexes the whole body of knowledge metadata for literature in a variety of formats and disciplines, with links. Estimates can be difficult at some point. However, Delgado López-Cózar et al. (2018) estimates size of Google Scholar at 331 million records in March 2017. There is also another study (Gusenbauer, 2019) that compares the amount of 12 academic search platforms and bibliographic warehouses. For example, there seems nearly 4 million records based on query from Web of Science (WoS) regarding Computer Science category. Top 10 sub-domains and their record counts can be seen in Figure 1.

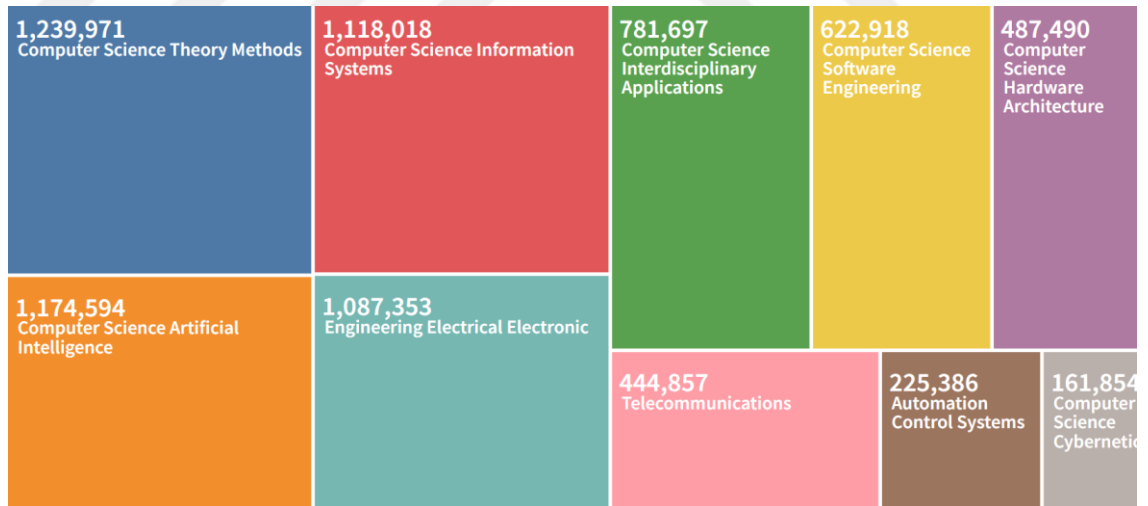


Figure 1. Distribution Of Computer Science Articles In WoS

Application areas regarding Computer Science are listed as Artificial Intelligence, Information Systems, Engineering Electrical Electronic, Interdisciplinary, Software Engineering, Telecommunications, Hardware Architecture, Automation Control Systems and Cybernetics.



In industry, government, and academia, science and technology roadmaps are used to depict the structural linkages between science, technology, and applications. In increasingly complex and uncertain contexts, roadmaps are used as decision aids to better coordination of operations and resources (Kostoff & Schaller, 2001). In an era of strategic science and high-investment initiatives, decision-makers want to be able to discover potential and attractive pathways and possibilities for technological emergence ahead of time in order to help them to choose the best path. Bioelectronics, nanotechnologies and blockchain are examples of new and developing science and technology that could have an impact on a wide range of industries. As a result, support systems are required to supplement regulation and policy instruments in order to achieve their goals — a variety of (and frequently shifting) industrial environments. This underlines the importance of timely and relevant strategic knowledge in order to make informed decisions and establish effective strategies (Robinson et al., 2013). At this point, analyzing past data is important to provide path regarding collaboration of science, technology and innovation.

Today, researchers have access to an increasing amount of digitized text. Online information from news resources and micro blogs, firm press releases, customer reviews of experiments and products, scientific publications and conversation, are examples of such writings. In fact, widespread digitalization activities across practically all industries will very surely increase the diversity and volume of unstructured data (Antons et al., 2020). The majority of academic writing is based on the idea that new research should build on and extend previous work — a practice that frequently means that any worthwhile academic enquiry begins with an investigation and analysis of all previous results on the subject at hand. At this point, a question reveals: Is it possible for scholars to navigate this massive maze of existing literature?

Thanks to text mining, data mining and Artificial Intelligence (AI), machine learning applications and techniques. They can provide algorithms, methods and tools to gather vast amounts of text and data from existing articles to be analyzed, classified, and sorted to find patterns and extract relevant information for a number of applications. Users can utilize text and data mining to detect and summarize previously unseen interconnections, as well as analyze and uncover patterns and relationships in a variety of databases where connections were previously difficult to make.

There is a new era regarding research in the 21st century. Rather than manual search for relevant academic publications, Technology examines and pulls significant scientific material from published literature in a methodical manner, classifies it according to numerous parameters, and connects the dots. In addition to, there can be data classification and analysis from a variety of dimensions and perspectives. Machine learning and deep learning are the operation point of creating new data models without having to program them explicitly. Data mining can be used to develop databases that can then be mined as well. Analysis of the past and present serves as a foundation for future forecasting.

There is also another popular word, “foresight” used as “Technological Foresighting” or “ForSTI” (Foresight for STI). “STI” stands here as Science, Technology and Innovation. In most dictionary definitions of foresight, two complementing qualities are mentioned. The first skill is foresight into possible future events, as well as the ability to make judgments about the consequences of existing patterns and situations. The second capability is caution, which entails being ready for the significant reactions that hard implications may necessitate. The combination of these qualities gives "foresight" a positive connotation that phrases like prediction, prophesy, and planning don't have. There is also another definition of foresight where Loveridge (2008) has described as institutional Foresight. This necessitates a concerted, built effort to implement, gather, and organize evidence-based opinions about forthcoming challenges and possibilities, as well as to use them as strategic information for decision-making. This is a lot more than a guessing game in which you try to figure out what will happen. Similarly, it is more than a planning effort geared at stating what should be done to be able to resolve specific problems or put specific resolutions in place (Miles et al., 2016). Foresight envisions an open and interdisciplinary debate and communication culture that facilitates the interchange of information between policymakers, industry, science, and society, fosters interaction, and supports networking and results implementation. In general, these studies use systemic, integrative techniques with a variety of instruments and methods to meet divergent interests and achieve consensus among all parties, necessitating the participation of important stakeholders (Meissner & Sokolov, 2013). STI investment may be a viable approach of reaching the goal of remaining competitive in today's globalized era (Hameed et al., 2016). Even at a micro level, innovation allows a company to develop competencies that are difficult to duplicate, resulting in a competitive advantage.

For innovations and new product development (NPD) approaches, customer involvement is either at the heart of the developments or their involvement is considered a supportive mechanism at different phases of various decision-making processes. Some academics believe that customers are the key for a technology to be accepted (Chang & Taylor, 2016), whereas some believe customer opinion is not so critical, especially for radical/discontinuous innovations or where there is an asymmetry between company and customer knowledge (Trott, 2001). There are a number of reasons for the negative opinion towards customer-centric decision support systems in NPD. Firstly, it is costly, especially if third party-based surveys and product analysis are considered. Secondly, it is believed that customers may not possess adequate information especially in high tech environments. Finally, due to the traditional customer-centric product analysis (i.e., surveys, focus groups, product trials), the gathered information is considered to be too static for fast-paced environments.

Innovative methods using social media data in customer-centric product analysis approaches minimize or eliminate the above-mentioned problems. Social-media data are used in a variety of situations from digital marketing to customer satisfaction analysis as social media-based operations are low-cost, dynamic, targeted and informative due to the new possibilities presented by the 4Vs of big data and advanced machine learning approaches (Saura, 2021). Many failed innovations also indicate the need for advanced product analysis systems.

In recent studies, social media data have been used for product analysis for different reasons, such as the examination of customer reaction to the launch of new products or technologies (Lipizzi et al., 2015; Nuortimo & Härkönen, 2018), the assessment of a product's competitive advantages (Liu et al., 2019; Jiang et al., 2019) and assistance with the development of next-generation products (Li et al., 2014; Hou et al., 2019; Mirtalaie et al., 2017). The majority of these studies build multi-stage models using different combinations of sentiment analysis, topic modelling, natural language processing, named entity recognition, and term extraction techniques to analyze user generated content (Saura & Bennett, 2019). These methods, however, have limitations in their ability to offer an end-to-end social media-based feedback mechanism for decision support systems in innovations and NPDs. Examining the literature specific to the product development and ideation steps, there are only a few studies in which social media data are used as a part of the NPD process.

## 1.2 Research Goal

Regarding the motivation, it is obvious that analyzing text from the point of text mining and data mining views is an important and hot topic. The volume, accessibility, and usefulness of unstructured data in the form of digital text are fast rising for study on STI. Combining text analysis with machine learning and deep learning overcomes the limitations of classical qualitative analysis in processing vast amounts of data. The main research goal in thesis study is to propose a novel approach/framework with combining/utilizing text analysis and deep neural networks regarding analysis of STI. There are two sub-research goals. The first part is performing opinion mining from social media for an innovative technology product via using multi-task deep neural networks. The second part is performing opinion forecasting for a selected technology domain based on estimation of word embedding matrix using Long Short-Term Memory (LSTM) networks.

For the first part, the major weakness in the literature is the design of the methodology and its linkage to the product development process. The current literature fails to offer strong examples of models that can retrieve the product development and innovation-oriented opinion of customers. Many studies in the literature are designed to retrieve customer sentiment regarding the product in general or existing features of the product (Hasson et al., 2019; Ibrahim & Wang, 2019; Liu et al., 2019). Studies where only sentiment analysis is performed for product analysis fail to provide detailed information to understand why the customers do or do not like the product. However, the studies that perform aspect-based analysis for feature-based product analysis do not offer innovative ideas that can be used for product development. There are only a few studies that aim to generate innovative ideas related to the specific features of the product via sentiment analysis of the customer comments shared on social media (Li et al., 2014; Mirtalaie et al., 2017). Such studies propose frameworks to identify sentiment and customer opinion to improve the next version of products or to introduce new types of products based on specific customer suggestions that focus on groups of product features.

Considering this notable gap in the literature, in this study we propose a framework that uses social media data to reveal the reasons for a failed innovative product from the customer perspective and to suggest new use cases and innovative ideas for product development. We selected the case of Google Glass as this product had a failed launch despite the technological offerings and the level of innovations it featured. Google Glass, a smart glass brand, is an optical

head-mounted device that can be controlled via the voice and motions of its user and can assist the user by displaying information on its screen. It is regarded as a useful case that offers rich data considering both positive and negative customer feedback. As our aim is to analyze product-related concerns and extract innovative ideas for product development rather than performing a quantitative customer satisfaction analysis, we integrated an opinion detection module into our system to clean the dataset by removing the user generated content which does not include useful feedback or suggestion but only expresses satisfaction/dissatisfaction about the reference product. To achieve our objectives, we propose the design of a multi-task deep neural networks (DNN)-based framework that learns sentiment analysis and opinion detection tasks jointly.

One of the best ways for businesses to communicate with customers is through social media platforms. Although social media-based decision support systems for the marketing industry are highly advanced, studies focused on product development and innovation are still in their infancy. The first section of the thesis presents a novel strategy that supports the study and development of products by using sentiment analysis and the opinion retrieval/mining theme. We present an end-to-end system for social media opinion retrieval that makes use of machine learning and Natural Language Processing (NLP) methods to accomplish this goal. The product Google Glass was picked as a use-case because, although having better technological capabilities, it failed to meet its sales goals. To train deep neural networks for emotion prediction and opinion detection tasks, we create a multi-task architecture. First, we categorize the tweets that contain specific helpful comments and recommendations into two groups based on their sentiment labels. While the positive and neutral tweets are utilized to glean creative ideas and find fresh applications for product development, the negative tweets are examined to find issues with the products themselves. The keyword clusters that we extract from each sentiment label group are displayed and interpreted. For the future generation of smart glasses, this work offers practical contributions in addition to methodological ones.

For the second part, by industrial, governmental, and academic actors, technological foresight and forecasting methods are used to understand and foresee technological changes with an aim of establishing scenarios, roadmaps and strategies (Kostoff & Schaller, 2001). Nations and organizations are monitoring the technological advancements to set strategies

and policies to establish potential pathways for the future. For this, it is essential to identify and monitor the technologies of future. Some scholars focused on the identification of emerging technologies (von Delft & Zhao, 2021; Pitt & McCarthy, 2021; Burmaoglu et al., 2019). Some scholars examined the technological developments in the recent past in specific areas and they developed potential scenarios or roadmaps (Nazarenko et al., 2022; Yun et al., 2021; Jissink et al., 2019; Park et al., 2020; Zou et al., 2020). Considering the technological forecasting area and especially the discussions of the foresight scholars in this field, we can observe a pendulum between quantitative and qualitative approaches. Especially, quantitative approaches in the last decade became even more popular thanks to text mining, artificial intelligence (AI) and machine learning technologies and techniques. Users can utilize these approaches to detect and summarize previously unseen interconnections, as well as analyze and uncover patterns and relationships in a variety of databases where connections were previously difficult to make.

Rotolo et al. (2015) creates a conceptual description of "emerging technologies" and connecting this work to the creation of a framework for the operationalization of technological emergence. The definition is created by combining a fundamental comprehension of the term, especially the concept of "emergence," with a survey of important innovation studies addressing definitional concerns of technological emergence. According to this study, radical originality, rapid expansion, coherence, significant influence, uncertainty, and ambiguity are all hallmarks of emergent technologies. Technological emergence indicators offer insightful information regarding setting R&D priorities. Porter et al. (2019) describes a working approach for determining topical word emergence scores from abstract record sets.

Having mentioned the efforts and methods trying to predict the future, technological foresight and forecasting is much more than a guessing game in which you try to figure out what will happen. Similarly, it is more than a planning effort geared at stating what should be done in order to solve specific problems or put specific solutions in place (Miles et al., 2016). Foresight envisions an open and interdisciplinary debate and communication culture that facilitates the interchange of information between policymakers, industry, science, and society, fosters interaction, and supports networking and results implementation. In general, these studies use systemic, integrative techniques with a variety of instruments and methods to meet divergent

interests and achieve consensus among all parties, necessitating the participation of important stakeholders (Meissner & Sokolov, 2013). However, the TF methods are an important piece and a tool in this complicated puzzle and it is essential to identify new ways of foreseeing the future. There was also a contest regarding measuring tech emergence (Porter et al., 2018).

Considering the relevant data science approaches in this field, the majority of them apply topic modelling or clustering methods (Zhang et al., 2017; Jebari et al., 2021; Qian et al., 2021; Behpour et al., 2021; Vahidnia et al., 2021; Ozansoy & Sağkaya, 2021; Delgosha et al., 2021; Qiu & Wang, 2020; Zeng, 2018; Li et al., 2019). In recent studies, focusing on more advanced approaches, deep learning methods also started to be utilized (Lu et al., 2021; Liang et al., 2021). From the point of utilizing word/document embedding features, there are also studies to evaluate the analysis of research topics and topic evolution (Vahidnia et al., 2021; Hu et al., 2019; Huang et al., 2020). Even though the data science-based methods are progressing in this field, there is no available TF approach where the future can be predicted based on the co-occurrence of terms. Any predictions in this field relies on linear trend analysis or statistical approaches to make actual predictions without expert opinion or interpretation.

Considering this gap in the literature, we are, for the first time, combining semantic word embedding features with deep learning methods to be treated as time-series data. We propose a novel framework that uses literature past embedding data in word level to predict evolving topics, clusters in future regarding emerging technology. In addition, we propose to use Word2vec in an online training fashion to obtain the monthly word embedding representations in an efficient way. The practical and methodological objectives of our study are as follows:

- Propose a general framework that analyses literature data, evaluates current topics/clusters and provides insight for future
- Utilize word embedding features with LSTM network as supervised regression task
- Propose a methodology for visualizing clustering with using cosine similarity metric
- Provide prediction for the future regarding cluster evolution

### **1.3 Thesis Organization**

The remainder of this study is structured as follows.

## **Chapter 2: Methods**

This chapter provides the details regarding the research methods in general. Firstly, investigation and analysis results are described from the point of systematic literature review. This sub-section is divided into two parts as opinion mining and opinion forecasting aspects. Secondly, selected/decided use cases and background information are talked. Thirdly, NLP area is analyzed, and applied methods and techniques are mentioned. Fourthly, text representation algorithms are provided. It consists of classical Bag of Words (BoW) methods and word embedding matrix styles. Lastly, deep learning algorithms and their mathematical and scientific background are reviewed.

## **Chapter 3: Models**

This chapter provides the details regarding the proposed models to achieving research goals. It is divided into two major sub-sections regarding opinion mining and opinion forecasting. Both sub-sections consist of Dataset, Pre-Processing, Feature Extraction, Model Creation, Visualization and Results steps. In Dataset, type of data is described, and collection method is talked. In Pre-Processing, NLP related methods, applications are studied on the gathered data. In Feature Extraction, details about representation of textual data are given prior to use in machine learning, deep learning models. In Model Creation, the proposed deep learning models are introduced. In Visualization, the followed methodology is mentioned to represent model output. In Results, both experimental and practical outcomes are given.

## **Chapter 4: Conclusions**

This chapter involves overall evaluation of the results and studies. It consists of Discussions, Contributions, Limitations and Future Works.



## Chapter 2

### Methods

#### 2.1 Systematic Literature Review

In this section, literature review is divided in two parts. In the first part, product analysis topic is reviewed from the point of opinion mining. Second part focuses on science, technology and innovation from the point of opinion forecasting view. Regarding systematic literature review several academic databases and resources (ACM Digital Library, IEEE, ScienceDirect, Scopus, SpringerLink, Taylor & Francis, Web of Science, Wiley Online Library etc.) are searched. The major works regarding this are satisfied with using Google Scholar. It provides a simple method for looking up scholarly literature in a wide sense. From one location, you may search academic publishers, professional societies, online repositories, universities, and other websites' papers, theses, books, abstracts, and court decisions. You can find pertinent research from scholarly publications worldwide by using Google Scholar. A sample screenshot from Google Scholar can be seen in the following figure. You are able to search articles from there with setting additional filters and save into your library or favorites.

**2.1.1 Opinion mining for analysis of product.** Rapid advancements in Internet of Services, Web 2.0, and social media have led to potential applications and better interactions with consumers. Product development and innovation activities can now be supported by knowledge retrieval applications. Consumers share their opinions in public domains, and their input can be used as the source of ideas for new generation products, services and business models. Two innovative approaches can be implemented to retrieve information from the crowd: 1) crowdsourcing models, where consumers are led to a certain task and information is retrieved (Djelassi & Decoopman, 2013; Schemmann et al., 2016), and 2) idea or opinion mining, where information is retrieved from generic public data with a specific approach (Li et al., 2014; Lipizzi et al., 2015). Both approaches may have advantages and disadvantages. Crowdsourcing is a better approach if the required information is scarce and lacking in detail, in cases where a company requires feedback for a certain stage of product development (i.e., prototyping). However, crowdsourcing-based approaches could be costly, and continuously implementing such

approaches is more difficult. For idea/opinion mining approaches, the optimal approach is to retrieve information from big data to ensure the continuity and efficiency of the process.

Table 1 summarizes the review of the relevant studies, considering their research focus and methodological approaches. Accordingly, the studies are grouped as: 1) company, product, or sector-specific customer feedback analysis using social media, 2) product review analysis using product-specific online marketplace or product forums, and 3) product review analysis for product development from social media data. The analysis of the literature studies summarized in Table 1 shows context and methodology-specific gaps and weaknesses. Most of the relevant studies performed a general sentimental analysis based on the sentiments of the content generated by the users without presenting specific feedback related to the product features. For example, a customer comment may consist of an overall positive sentiment (i.e., “I love this product”) but may lack product- or feature-specific useful feedback (i.e. “I love this product’s camera quality”). For product development purposes, it is more valuable to identify customer feedback that has an opinion to improve the current offering or has suggestions for new types of products. Table 1 shows that there are some studies that utilize social media data for product development. These studies focus on finding new features that can be integrated to the next version of the products. However, to the extent that we know, there is no prior study that proposes an end-to-end framework that uses social media data to reveal the reasons for a failed innovative product from the customer perspective and to suggest new use cases and innovative ideas for product development. This gap in the literature can be addressed by implementing sentiment analysis and opinion detection modules together to identify customer comments with product-specific feedback and suggestions. Therefore, in this study we focus on developing a multi-task learning-based opinion retrieval system that can analyze social media data for product analysis and development.

Table 1  
Summary of the Related Works

Group of Studies	Authors	Research Aims	Methods	In comparison with the previous studies
Company, product or sector specific customer feedback analysis using social media	Hasson et al. (2019) Rane and Kumar (2018) Botchway et al. (2019) Ibrahim and Wang (2019) Mai and Le (2020) Lipizzi et al. (2015)	To perform sentiment analysis and determine the most common customer complaints  To identify customers' primary topics of concern  To compare customer satisfaction surveys with social media based customer feedback  To perform sentiment analysis on tweets related to a specific company and determine the most popular hashtags about the products and brands of the related company  To analyze early reactions of customers for new products	Supervised learning techniques combined with natural language processing, topic modelling and network analysis modules	Do not offer a detailed sentiment-based analysis for product development with a word network that can be utilized by decision makers but mostly present only sentiment category count related to each product/service of the company  Do not have an opinion detection module to clean the dataset by removing the entries that do not contain useful feedback  Mostly focus on only analyzing the negative tweets to determine the products and services that are perceived negatively by the customers
Product review analysis using a product specific online marketplace, or product forums	Liu et al. (2019) Sun et al. (2019) Basiri et al. (2020), Eldin et al. (2020), Jiang et al. (2019)	To analyze product specific user comments retrieved from a product specific platform  To assess the competing products from the customer perspective  To identify the degree of a review's informativeness	Traditional supervised learning and deep learning techniques combined with domain specific lexicon generation approaches and named entity recognition methods	Less general solutions are offered as customer feedback or opinion retrieved from product or service specific online marketplace or product forums are analyzed  Mostly domain specific features such as membership status of the user, availability of a public user image, etc. are used in addition to the review text  Do not offer innovation-oriented ideas for product development  Do not offer an automated end to end approach from data retrieval to the extraction of word network of customer feedback that can be utilized for specific product development

Product review analysis for product development from social media data	Mirtalaie et al. (2017), Li et al. (2014)	To propose a decision support framework to retrieve product specific innovative ideas from social media	Similarity metrics, rule based approach or a knowledge base for sentiment analysis, word clouds, various metrics to evaluate the reviews (e.g. influence score, expertise score, review rating)	Machine learning/deep learning techniques are not used  Performs cross-domain analysis to extract innovative ideas for the reference product
--	---	---	---	--

Previous studies illustrate the application of text mining in different phases of decision-making in product and technology development. From these, it is evident that social media data are used for to examine emerging technologies, product competitiveness or life cycle analysis and overall opinion gathering. As the increasing amount of social media data provides opportunities for building decision support systems regarding product or service improvement, some of these studies focus on processing the social media content for product-related feedback analysis (Botchway et al., 2019; Hasson et al., 2019; Jiang et al., 2011; Mirtalaie et al., 2017; Saura & Bennett, 2019).

Micro-blogging services are one of the main data sources for companies to gather feedback from potential customers and improve their services and products in the light of this analysis (Araque et al., 2017). Twitter is an extremely popular micro-blogging service and has been used widely by academics and companies as an important source of customer feedback. Hasson et al. (2019) concluded that Twitter data is a less costly alternative to customer satisfaction surveys. They used Twitter data belonging to a large biotechnology company and compared the customer feedback gathered from social media with the survey responses in terms of content and value. For this purpose, they first applied pre-processing operations on the Twitter data, used another already labelled sentiment dataset to train a classifier, applied the obtained sentiment analysis model on their original dataset to classify each tweet with a sentiment label, and constructed a hash-tag co-occurrence matrix for the top six products and/or services of the company. Finally, they retrieved a sample of survey data from the company and comparatively analyzed the sentiment analysis and survey results. According to the research, several products/services that are not the focus of the company were discussed more than some of their ground-breaking work in both negative and positive tweets, showing that the company and customers may differ in their interests in the products and services. Additionally, their findings indicated that, in comparison to customer satisfaction surveys, analysis of social media material can offer more dynamic data and fresh perspectives on consumer happiness.

In a recent study, Ibrahim and Wang (2019) analyzed Twitter data related to five leading UK online retailers by combining topic modelling, sentiment analysis and network analysis techniques, with the aim of identifying the main concerns of the customers. Different from our study, they did not experiment with different DNN based models for sentiment analysis but instead used a sentiment analysis tool to identify the negative tweets. They then specifically highlighted the services and products perceived negatively by their customers. They used Latent Dirichlet Allocation (LDA) for topic modelling and listed the most important eight topics from the negative tweets of the customers. Rathan et al. (2018) proposed a Twitter-specific sentiment analysis model using features such as emoji detection, emoticon detection and spelling correction. The model was applied to the “Smartphone” domain. The authors presented an aspect-based analysis considering different attributes of smartphones such as battery, camera and display.

Liu et al. (2019) applied, similar to our study, a two-step supervised learning approach to analyze social media text data. The main goal of the study was to assess competing products from the perspective of customers. The first step of their approach was sentiment analysis performed using a domain-specific sentiment lexicon. They subsequently built a classification model to detect the comparative user-generated content. Finally, they presented the most important advantages of the target product compared to its competitors. The results indicated that sentiment analysis plays a key role in analyzing customer feedback from different perspectives. Sun et al. (2019) proposed a machine learning-based framework to identify the degree of a review’s informativeness, using data from an online electronic marketplace in China. As stated in this study, sentiment analysis of user-generated content has an important role in determining the informativeness of a review. Therefore, in addition to the binary classification model built to determine whether the tweet includes a feedback or suggestion, we incorporate sentiment analysis into our framework to analyze positive and negative reviews for aspect-based opinion retrieval and analysis.

In another study, Rane and Kumar (2018) analyzed 14,640 tweets associated with six major US airlines. They first manually labelled all tweets. Subsequently, after text pre-processing operations, they used a word embedding technique, Doc2vec, and seven different classifiers to classify the tweets as negative, neutral or positive. They presented a comparative analysis of the classifiers’ performances and identified the common terms which appeared in

the negative feedback of the customers. Similarly, Botchway et al. (2019) applied sentiment analysis techniques to Twitter data to analyze the customer feedback related to the products and services of one of the largest banks in Europe. They used a tool for sentiment analysis based on rules designed to process social media data. They also presented top hashtags which appeared in the customers' tweets. Basiri et al. (2020) performed sentiment analysis in the medical domain by applying various deep learning models to the drug reviews shared in Drugs.com. They compared their proposed deep fusion models with the existing traditional and deep learning-based techniques.

Mai and Le (2020) used deep learning techniques to examine user comments regarding smart-phone products shared on a social media platform. Similar to our study, they built a multi-task learning framework to analyze user comments. Specifically, while we perform multi-task learning for opinion detection and sentiment analysis tasks, they train sentence-level and aspect-level sentiment analysis tasks jointly to design a more generalizable sentiment analysis model. Different from our study, their framework presents the proportion of positive and negative comments for specific attributes of a product rather than generating innovative ideas and new use cases. Similar to the main goal of our study, Mirtalaie et al. (2017) aimed to generate innovative ideas from user reviews retrieved from social media. They proposed a framework consisting of three stages. The proposed framework identifies related products with the reference product with a cross-domain analysis and identifies new features that can be integrated into the future versions of the reference product. The studies applying sentiment analysis in various domains show that is a key technique for building managerial decision-making tools for product and service improvement.

Having reviewed the literature and demonstrated the relevance of our study in Table 1, it is evident that there are limited studies where such approaches are being implemented for the purpose of retrieving feedback and opinions for product development or innovation-oriented processes. Most studies are limited to offering direct solutions that can be used for product analysis or development but mostly present an overall quantitative assessment about the reference product. There are also methodological gaps in terms of combining deep learning and advanced word embedding techniques with the aim of retrieving useful product and service-related suggestions and feedback by benefiting from user generated content in social media.

**2.1.2 Technological forecasting for technology and innovation.** Academia, research institutions, think tanks, and business intelligence agencies are following, analyzing, and anticipating developing innovations in response to increased competition in new industries and markets. It is critical to examine technological frontiers and their future developments for providing funding for research and development (Ren & Zhao, 2021).

Table 2 summarizes the review of the relevant studies, considering their research focus and methodological approaches. Accordingly, the studies are grouped as: 1) Trend & Topic Analysis in Science, Technology and Innovation and, 2) Technology Forecasting & Foresighting. The analysis of the literature studies summarized in Table 2 shows context and methodology-specific gaps and weaknesses.

In the first group, “Trend & Topic Analysis in Science, Technology and Innovation” as shown in Table 2, a majority of the relevant studies performed topic modelling methodology based on features extracted/created from either term frequency or word embedding. Zhang et al. (2017) used the journal Knowledge-based Systems as source of text data to perform bibliometric analysis from 1991 to 2016. Artificial intelligence-supported e-learning domain is selected to be reviewed as trend analysis regarding timeframe 1998-2019 (Tang et al., 2021). A large-scale analysis of the transformation of biomedical and life sciences is performed with using the citation contexts based on collected papers published in PubMed Central between 2008 and 2018 (Jebari et al., 2021). A similar medical domain is worked via processing electronic health records (Qian et al., 2021). In this study, bibliometric data is retrieved from Web of Science and lda2vec is mainly applied as unsupervised learning. Clustering strategy with using LDA followed by SVD is applied for finance domain journal abstracts from 1974 to 2020 (Behpour et al., 2021). They introduced a weighted temporal feature to emphasize time factor during topic clustering. Mun et al. (2021) introduces function score to determine the relative significance of a feature in a technological field at a particular period. For emphasizing time concern, Lu et al. (2021) utilized Long Short-Term Memory as deep learning model with using author defined keyword frequency. In another study, LSTM is utilized again with combining auto regression models (Liang et al., 2021). Term frequency is considered rather than word embedding or document embedding. Traditional clustering algorithms and topic modelling are used on embedding feature sets Doc2Vec, FastText, BERT generated from title and abstract in the articles (Vahidnia et al., 2021). Word2Vec model is also popular as feature

representation method in such studies. For example, Hu et al. (2019) implemented an approach with combining Google Word2Vec model spatial autocorrelation analysis to evaluate topic evolution of scientific literatures. In a similar study, Word2Vec model is combined with cluster percolation algorithm to identify research themes (Huang et al., 2020). As a novelty, author keyword is fed separately from abstract and title while generating word embeddings. For electronic business research domain, a predictive analysis is performed with using term frequency as an input to LDA based on collected article reviews between 1994–2020 (Ozansoy and Sağkaya, 2021). Emerging COVID-19 research trends are evaluated with using co-word analysis (Verma and Gustafsson, 2020). Delgosha et al. (2021) combined topic modelling technique LDA and thematic analysis regarding discovering IoT implications. Through topic modeling approaches like LDA with BoW, Gupta et al. (2022) aims to identify new directions, paradigms as predictors, and associations between each topic. For this purpose, a 30-year period-long empirical examination of 3269 research publications from the Journal of Applied Intelligence was conducted.

In the second group, Technology Forecasting & Foresighting, majority of the relevant studies do not follow traditional machine learning, deep learning and topic modelling techniques. Liu et al. (2021) applied text mining methods and social network analysis as focusing on gene editing patents. Text mining and f-term analysis are applied to discover emerging technology opportunities (Song et al., 2017). In a similar study, text mining methods are combined with expert review process as focusing on retail industry (Ozcan et al., 2021). For robotics domain, patent abstract data is utilized by LDA as topic modelling with integrating semantic similarity (Qiu and Wang, 2020). Kim and Bae (2017) also worked on patent data to forecast promising technology with introducing cooperative patent classification. In another study, technology fields in patents are evaluated with collaboration of network analysis and multilateral brokerage analysis (Huang and Su, 2019). For healthcare domain, two technological forecasting methods, network analysis and expert review are utilized (Lee et al., 2019). Zeng (2018) benefits from online communities to foresight renewable energies with using LDA technique. Combining clustering with expert review is applied for the use case of perovskite solar cell technology as a forecasting technology trend study (Li et al., 2019). 13 910 utility patents issued between 2000 and 2014 are used in a case study here (Chen et al., 2017). They combined patent theme analysis and trend analysis. In order to create data-driven technology roadmaps, Kim and Geum (2021) offers a practical methodology. The proposed



architecture is divided into three stages: Layer mapping, content mapping, and opportunity mapping. Layer mapping is the initial stage, where subject modelling is used to determine the sub-layers of the technology roadmap. In the second stage, the keyword network analysis is used to map the contents. In the last stage, with the use of link prediction, opportunity discovery is done to foresee potential future innovation opportunities. Zhou et al. (2020) employs GAN to get around the issue of not having enough training data. After then, they apply deep neural network classifier. Denter et al. (2022) suggests a new method for locating promising patents. They used machine learning algorithms to integrate link prediction with textual patent data for camera technology domain. The following six perspectives—key, outlier, unoccupied, emerging, new, and converging technologies—along with their pertinent patent analysis methodologies are investigated in (Jee et al., 2021) with the goal of creating a taxonomy of promising technologies. Their conclusions show that the definition of a promising technology might vary depending on one's point of view, and this can serve as a foundation for further research.

Table 2  
*Summary of the Related Works*

Group of Studies	References Summary of Studying Methods	Research Aims	In comparison with the previous studies
Trend & Topic Analysis in Science, Technology and Innovation	<ul style="list-style-type: none"> <li>Topic Modelling <ul style="list-style-type: none"> <li>○ Zhang et al. (2017)</li> <li>○ Jebbari et al. (2021)</li> <li>○ Qian et al. (2021)</li> <li>○ Behpour et al. (2021)</li> <li>○ Vahidnia et al. (2021)</li> <li>○ Ozansoy and Sağkaya (2021)</li> <li>○ Delgosha et al. (2021)</li> <li>○ Gupta et al. (2022)</li> </ul> </li> <li>Social Network Analysis <ul style="list-style-type: none"> <li>○ Qian et al. (2021)</li> </ul> </li> <li>Machine/Deep Learning <ul style="list-style-type: none"> <li>○ Lu et al. (2021)</li> <li>○ Liang et al. (2021)</li> </ul> </li> <li>Traditional Clustering Algorithms <ul style="list-style-type: none"> <li>○ Vahidnia et al. (2021)</li> </ul> </li> <li>Document and Word Embedding <ul style="list-style-type: none"> <li>○ Vahidnia et al. (2021)</li> <li>○ Hu et al. (2019)</li> <li>○ Huang et al. (2020)</li> </ul> </li> <li>Bibliometrics <ul style="list-style-type: none"> <li>○ Ozansoy and Sağkaya (2021)</li> <li>○ Verma and Gustafsson (2020)</li> </ul> </li> <li>Miscellaneous <ul style="list-style-type: none"> <li>○ Tang et al. (2021)</li> <li>○ Mun et al. (2021)</li> <li>○ Small et al. (2014)</li> </ul> </li> </ul>	<p>To detect and predict topic change in Knowledge-based Systems</p> <p>To identify trends in artificial intelligence-supported e-learning</p> <p>To detect the evolution of research topics</p> <p>To explore hot topics and trends of Electronic Health Records Literature</p> <p>To detect trend over a set of finance journals</p> <p>To detect research topic trends.</p> <p>To detect and extract research topics from academic documents.</p> <p>To predict emerging research topics.</p> <p>To understand the topic evolution of scientific literatures</p> <p>To discover overlapping community for identifying research themes.</p> <p>To understand the patterns in a technological domain's fundamental development.</p>	<p>Mostly focus on term frequency to create co-occurrence map</p> <p>Mostly focus co-citation network.</p> <p>Author keyword and Title, Abstract parts are evaluated separately.</p> <p>From the point of text processing, mostly focus on Abstract and use TF-IDF for performing topic modelling.</p> <p>Author Keyword is identified as main input and combined with community indicators.</p> <p>Do not include keywords in the articles and mostly focus on title and abstract.</p> <p>Rather than embedding of word or documents, frequency is preferred.</p> <p>Global Word2Vec training is applied rather than online/local training.</p> <p>Author Keyword is fed separately from abstract and title while generating word embeddings.</p> <p>Mostly focus on technical verbs as 1-Gram with machine learning-based extraction technique.</p>

		<p>To analyse electronic business research from the point of historical and predictive aspects.</p> <p>To investigate the new COVID-19 research trends in the area of management and business.</p> <p>To discover IoT implications.</p>	<p>Mostly focus on term frequency to feed LDA.</p> <p>Traditional feature representation technique is used for co-word analysis.</p> <p>Traditional feature representation technique, bag-of-words is used for input to LDA.</p>
Technology Forecasting & Foresighting	<ul style="list-style-type: none"> <li>• Bag of Words <ul style="list-style-type: none"> <li>○ Liu et al. (2021)</li> <li>○ Ozcan et al. (2021)</li> <li>○ Lee et al. (2019)</li> </ul> </li> <li>• Social Network Analysis <ul style="list-style-type: none"> <li>○ Liu et al. (2021)</li> </ul> </li> <li>• Expert Review <ul style="list-style-type: none"> <li>○ Ozcan et al. (2021)</li> <li>○ Lee et al. (2019)</li> <li>○ Li et al. (2019)</li> </ul> </li> <li>• Topic Modelling <ul style="list-style-type: none"> <li>○ Qiu and Wang (2020)</li> <li>○ Zeng (2018)</li> <li>○ Li et al. (2019)</li> <li>○ Chen et al. (2017)</li> <li>○ Kim and Geum (2021)</li> </ul> </li> <li>• F-Term Analysis <ul style="list-style-type: none"> <li>○ Song et al. (2017)</li> </ul> </li> <li>• Network Analysis <ul style="list-style-type: none"> <li>○ Huang and Su (2019)</li> <li>○ Lee et al. (2019)</li> </ul> </li> <li>• Machine/Deep Learning <ul style="list-style-type: none"> <li>○ Zhou et al. (2020)</li> <li>○ Denter et al. (2022)</li> </ul> </li> <li>• Miscellaneous <ul style="list-style-type: none"> <li>○ Kim and Bae (2017)</li> <li>○ Bengisu and Nekhili (2006)</li> <li>○ Jee et al. (2021)</li> </ul> </li> </ul>	<p>To forecast technology focusing on gene editing patents.</p> <p>To identify technology road mapping for retail industry.</p> <p>To forecast technology focusing on robotics domain.</p> <p>To forecast promising technology through patent analysis</p> <p>To discover new technology opportunities based on patents.</p> <p>To identify innovative aspects of technological interdisciplinarity.</p> <p>To foresight promising technologies for healthcare</p> <p>To foresight renewable energies</p> <p>To anticipate technological developments in the field of perovskite solar cells</p> <p>To forecast emerging technologies through science and technology databases</p> <p>To develop data-driven technology roadmaps</p>	<p>Do not apply traditional machine learning, deep learning and topic modelling techniques.</p> <p>Mostly focus on term frequency and TF-IDF during text mining process.</p> <p>Patent abstract data is used for feeding LDA.</p> <p>Do not apply traditional machine learning, deep learning and topic modelling techniques.</p> <p>Do not apply traditional text mining techniques.</p> <p>Do not apply traditional machine learning and deep learning techniques.</p> <p>Combined network analysis with expert panel/review Focus on LDA technique</p> <p>Combined clustering with expert review</p> <p>Combined patent theme analysis and trend analysis</p> <p>Integrated topic modelling and link prediction</p> <p>Applied data augmentation and deep neural networks</p>

## 2.2 Case Study

In this section, corresponding case studies regarding opinion mining and opinion forecasting are described.

### 2.2.1 Opinion mining. For opinion mining part, analysis of a technological product,

Google Glass is selected as a use case. Google Glass is a head-mounted display wearable computer that looks like glasses. It acts as a hands-free smartphone that allows users to access the mobile internet browser, camera, maps, calendar, and other apps via voice commands. There are numerous literature studies regarding use cases for Google Glass (Wei et al., 2018; Carrera et al., 2019; Berger et al., 2017). Google Glass is an innovative product, and this use case aims to retrieve opinion of community from social media, micro-blogging environment Twitter.

**2.2.2 Technological forecasting.** For opinion forecasting part, analysis of a common topic for STI, Text Mining is selected as a use case. Text Mining is a sub-domain under Computer Science, Artificial Intelligence research area. Text mining is the process of transforming unstructured text into a structured format so that it can be used in other applications to discover fresh insights and meaningful patterns. There are numerous literature studies regarding utilization of text mining (Pejic-Bach et al., 2020; Hao et al., 2018; Nie & Sun, 2017). This use case aims to predict and forecast evolvments of topics, clusters regarding text mining subject via analysis of academic, literature data.

## **2.3 Natural Language Processing**

NLP is a common automatic computational processing of human languages is referred to as this phrase. This comprises algorithms that accept human-created text as input as well as algorithms that generate natural-looking text outputs. (Goldberg, 2017). A sample representation of human language activation is visualized in the following figure.

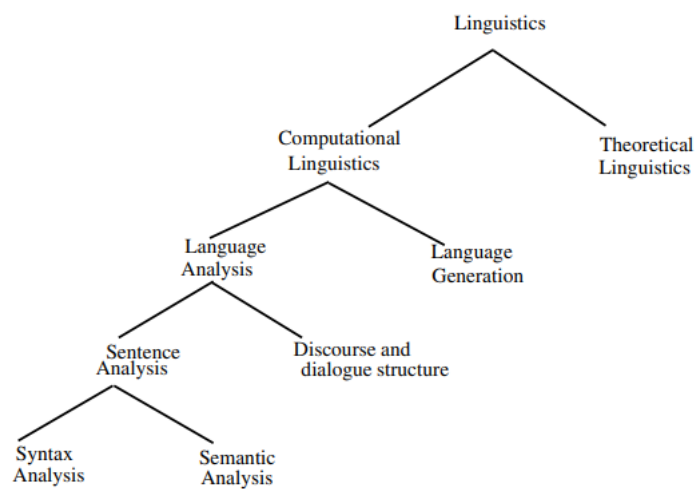
Machines must have a strong comprehension of natural language in order to do automatic text analysis. However, machines with this capability are still a long way off. Because natural languages are often large and sophisticated, it is difficult to create a program that can grasp them is needed, with an endless number of variants in sentences. Furthermore, Natural languages contain ambiguity, as many words have many meanings. When employed in different settings, the same sentence might have different meanings. As a result, creating programs that accurately interpret the natural language is a difficult undertaking (Chowdhary, 2020).

Some of the well-known NLP applications can be listed as in the following

- Text generations/dialogues
- Knowledge acquisition
- Question-answering (QA)

- Automatic summarization of texts
- Automatic language translation
- Information extraction
- Text classification into categories
- Retrieval of information
- Processing large texts from the point of searching and indexing

Components of NLP can be seen in Figure 2. In this thesis study, mainly we focus on Sentence Analysis and Semantic Analysis components.



*Figure 2. Components Of NLP (Chowdhary, 2020)*

NLP allows computers to comprehend natural language in a similar manner to how humans do. Natural language processing employs both spoken and written language, AI methods/models to take real-world data, interpret it, and make sense of it in a computer-readable format. The input is transformed to computer-readable code at some point throughout the processing.

There are two main phases about NLP: data pre-processing and algorithm development. Pre-processing includes preparing text data and cleaning it for computers to be able to analyze it. This stage can include following methods: Tokenization, stop words removal, Lemmatization, Stemming, POS (Part-of-Speech) tagging, Named Entity Recognition (NER), lexicon analysis etc. After this stage, it requires an algorithm to process the prepared and cleaned data. In that second stage, rules-based system or machine learning-based system can be used for this purpose. In this study, it was benefited from NLP stages and techniques to be

able to use textual data with machine learning, deep learning algorithms.

There are several literatures works regarding utilizing of NLP domain. To reveal numerous difficulties linked to COVID-19 from public viewpoints, Jelodar et al. (2020) employed an NLP technique based on topic modeling and automated extraction of COVID-19-related comments from social media. They also look into how to use an LSTM recurrent neural network to classify COVID-19 comments by sentiment. Kan et al. (2020) examine papers that use NLP as their primary analytical technique to show how textual data can be used to advance management ideas across many fields. They go over the many toolkits and procedures for using NLP as an analytical technique, as well as its benefits and drawbacks. Tenney et al. (2019) look at BERT and try to figure out where linguistic data is stored in the network. They discovered that the model accurately depicts the processes of the typical NLP pipeline, with the regions responsible for each step appearing in the expected order: POS tagging, parsing, NER, semantic roles, and coreference. Alshemali and Kalita (2020) explore how adversarial examples are currently being used to trick DNNs and provide a complete overview of how they might be used to increase the robustness of DNNs in NLP applications. They present a taxonomy for categorizing adversarial writings in this study, which summarizes existing ways for generating them.

## 2.4 Text Representation Algorithms

In this section, text representation algorithms used in this study are described in two sub-sections. Text representation is needed to convert text into a structured form to be processed by the DNN architectures. In first sub-section, it is talked about Bag of Words models. In the second sub-section, word embeddings are mentioned.

**2.4.1 Bag of words.** The traditional BoW techniques used as baseline methods in our study are based on TF and TF-Inverse Document Frequency (TF-IDF) metrics (Skansi, 2018). In BoW representation, the text is described in terms of the word occurrences of each word in the vocabulary. We used TF and TF-IDF metrics to represent the occurrence of each word. In TF-based representation, each word is represented with its frequency in the given text. TF is formulated as in the following way where  $t$  stands for term and  $d$  is document:

$$TF(t, d) = (\text{count of } t \text{ in } d) / (\text{number of words in } d) \quad (1)$$

Document Frequency (DF) finds the relevance of a document in a corpus of documents, Similar to TF in many ways. The primary distinction is that DF employs the quantity of occurrences of term  $t$  in the document set  $N$ , whereas TF applied a frequency counter for a term  $t$  in document  $d$ .

IDF, which quantifies the informativeness of a term  $t$ , is the inverse of document frequency. When we compute IDF, we will see that it is quite low for the most frequently occurring words, such as stop words (since stop words like "is" appear in practically every document):

$$IDF(t) = \log\left(\frac{N}{DF}\right) \quad (2)$$

TF-IDF-based representation not only uses the frequency of the term in the given text but also to what extent that word is common in the entire corpus. It is formulated as in the following way:

$$TF - IDF(t, d) = TF(t, d) * \log\left(\frac{N}{DF}\right) \quad (3)$$

Let's have an example for following documents and calculate TF and TF-IDF values.

Document 1: "This is first document"

Document 2: "This document is second document"

Document 3: "There is no document"

According to above documents, vocabulary consists of "this", "is", "first", "document", "second", "there", "no". Let's calculate TF values based on above formula in

Table 3.

Table 3

*Calculation of TF Values*

Term	Document 1	Document 2	Document 3
this	$\frac{1}{4}=0.25$	$\frac{1}{5}=0.2$	0
is	$\frac{1}{4}=0.25$	$\frac{1}{5}=0.2$	$\frac{1}{4}=0.25$
first	$\frac{1}{4}=0.25$	0	0
document	$\frac{1}{4}=0.25$	$\frac{2}{5}=0.4$	$\frac{1}{4}=0.25$

second	0	$1/5=0.2$	0
there	0	0	$1/4=0.25$
no	0	0	$1/4=0.25$

Now, let's calculate IDF values in Table 4.

Table 4  
*Calculation of IDF Values*

Term	IDF
this	$\text{Log}(3/2)=0.18$
is	$\text{Log}(3/3)=0$
first	$\text{Log}(3/1)=0.48$
document	$\text{Log}(3/3)=0$
second	$\text{Log}(3/1)=0.48$
there	$\text{Log}(3/1)=0.48$
no	$\text{Log}(3/1)=0.48$

And, finally let's compute TF-IDF values in Table 5.

Table 5  
*Calculation of TF-IDF Values*

Document	this	is	first	document	second	there	no
Document 1	0.045	0	0.12	0	0	0	0
Document 2	0.036	0	0	0	0.096	0	0
Document 3	0	0	0	0	0	0.096	0.096

There are several literatures works regarding utilizing of Bag of Words approach. HaCohen-Kerner et al. (2020) measures the influence of using BoW technique for text classification task. Cummins et al. (2018) utilizes training data from an additional dataset to investigate various BoW paradigms for sentiment identification. Silva et al. (2018) applies BoW

via encoding the local structures of a digital item in graphs. BoW feature engineering strategy is applied for assessing movement data and distinguishing between pathology patients and healthy people (Rastegari & Ali, 2020). BoW paradigm offers many customization options for your unique text data and is easy to understand and use. However, there are shortcomings regarding this technique. The vocabulary needs to be carefully planned, especially to control the size, which affects how sparsely the text represents things. Because the models must use so little data in such a large representational space, sparse representations are more challenging to describe for both computational (space and temporal complexity) and informational reasons. In addition to, the main drawback of BoW representation is that it disregards the semantic meaning and word orders and hence poorly represents the text.

**2.4.2 Word embeddings.** Distributional semantics is the categorization and quantification of semantic similarities among linguistic elements based on their distribution in the use of a language. For distributional semantics, vector space models have long been utilized to represent text documents and queries as vectors. There are different NLP algorithms benefit from the employing vector space models to express words in an N-dimensional space because in the new vector space, it results in clusters of text that are comparable (Goyal et al., 2018). Bengio et al. (2000) invented the term word embedding in their study "A Neural Probabilistic Language Model."

Word embedding is an advanced text representation technique where words from the vocabulary are converted to vectors of real values in a low-dimensional space, related to the size of the original vocabulary. This representation technique can capture the semantic and syntactic similarity between the words, and hence can represent their context text better (Goyal, 2018). Word embedding is a common NLP technique that involves mapping "words or sentences from the vocabulary to vectors of exact values." In terms of concept, it entails mathematical embedding from a one-dimensional space into a continuous vector space with one dimension per word with a substantially smaller size. These models are generally designed based on the use of neural network architectures.

Word embedding models have shown to be more efficient than the BoW models or one-hot-encoding systems that were previously utilized, which were made up of sparse vectors of the same size as the vocabulary. The vastness of the vocabulary and the labeling of the word or



document in it at the index location caused the vectoral representation to be sparse. This opinion has been replaced by word embedding, which uses the surrounding words of all individual words to extract information from the text and feed it to the model. As a result, embedding is now possible as a dense vector, which reflects the projection of individual words in a continuous vector space. Thus, embedding refers to the coordinates of the object.

Word embeddings (Baroni et al., 2014; Li et al., 2015) are generally divided into two sorts based on the methods used to create them. “Prediction-based” approaches are similar to neural language models where they make use of local data (for example, the context of a word). “Count-based” models, alternatively, are methods that utilize global information, such as corpus-wide data like word counts and frequencies (Almeida and Xexéo, 2019).

There are several implementations and approaches regarding predictions-based models. Bengio et al. (2003) introduced derived embeddings as a result of neural network training. Neural Networks with Hierarchical Softmax architecture are introduced by Morin and Bengio (2005). Mnih and Hinton (2007) introduced the log-bilinear mode. A multi-task neural network is trained with using both unsupervised and supervised data (Collobert & Weston, 2008). The best resulting type of Word2Vec was introduced based on Log-linear Model and Negative Sampling (Mikolov et al., 2013). Bojanowski et al. (2017) trained the embeddings at the n-gram level, in order to provide generalization for not seen data.

There are also several implementations and approaches regarding count-based models. Deerwester et al. (1990) introduced Latent Semantic Analysis (LSA) where SVD is performed on a term-document matrix. All of the corpus is scanned at a time and a word-word co-occurrence matrix is built (Lund and Burgess, 1996). Rohde et al. (2006) used standardization methods to prevent overuse of very common phrases cooccurrence counts. Usage of Canonical Correlation Analysis (CCA) between left and right circumstances to inspire word embeddings is introduced. (Dhillon et al., 2011). Lebrete and Collobert (2013) applied a modified version of Principal Component Analysis (PCA) to the term-context matrix. Pennington et al. (2014) introduced GloVe. In this study, we used two word embedding techniques, Word2Vec and GloVe, to represent the tweets in our dataset.

There are several literatures works regarding utilizing of Word Embeddings approach.

For sentiment categorization in social media, Wang et al. (2018) study the effects of word embedding and LSTM. Word embedding models are used to turn words in postings into vectors. The word sequence in phrases is then sent into an LSTM, which learns long-distance contextual dependency between words. Mohd et al. (2020) presents an automatic summarizer that uses the distributional semantic model to capture semantics and provide high-quality summaries utilizing word embeddings. They tested the proposed summarizer on different datasets and compared the results to other state-of-the-art summarizers. With using word embeddings, Roy et al. (2018) investigates the impact of changing the following two parameters: term normalization and training collection selection on ad hoc retrieval performance. They give quantitative estimates of word vector similarity derived under various conditions, as well as a query expansion job based on embeddings, to better understand the effects of these parameters on effectiveness of information retrieval. Meng et al. (2019) present a spherical generative model for concurrently learning unsupervised word and paragraph embeddings. They present an effective optimization approach to learn text embeddings in spherical space with guaranteed convergence. The approach is highly efficient and provides cutting-edge results on a variety of text embedding tasks, such as word similarity and document clustering.

In Figure 3, a sample word embedding data is visualized regarding applying PCA with the help of Embedding Projector tool (Smilkov et al., 2016; Abadi et al., 2019).



Figure 3. A PCA Projection Of Word Embedding Data

We are also able to find nearest neighbor in that embedding projector tool as shown in Figure 4. For example, nearest neighbor words for “health” word are listed there according to cosine distance.



Figure 4. Nearest Neighbor Points To Label/Word “Health”

The Word2Vec model is one of most utilized and transformed models in various NLP tasks (Ergen & Kozat, 2017; Kim et al., 2020; Li & Shah, 2017; Ren et al., 2019). Word2vec models' vector representations of words have been proved to have semantic content and are beneficial in a variety of NLP tasks. The main aim in this method is to learn the word associations or vector representation of the words in a given set of texts. In the Word2Vec model, with the use of a two-layered neural network architecture, the words that have similar semantic meanings are expected to have word vectors with high similarity. The Word2Vec approach can be implemented using one of the two model architectures which are known as Continuous BoW (CBoW) and skip-gram. Whereas in CBoW-based architecture the current word is predicted from a window of wraparound context words, in skip-gram-based architecture the surrounding window of context words are forecasted from the current word. The main advantage of skip-gram is to incorporate the order of context words into the word embedding process, hence resulting in the better representation of infrequent words (Bhoir et al., 2017). CBoW and Skip-Gram representations can be seen in Figure 5 and Figure 6.

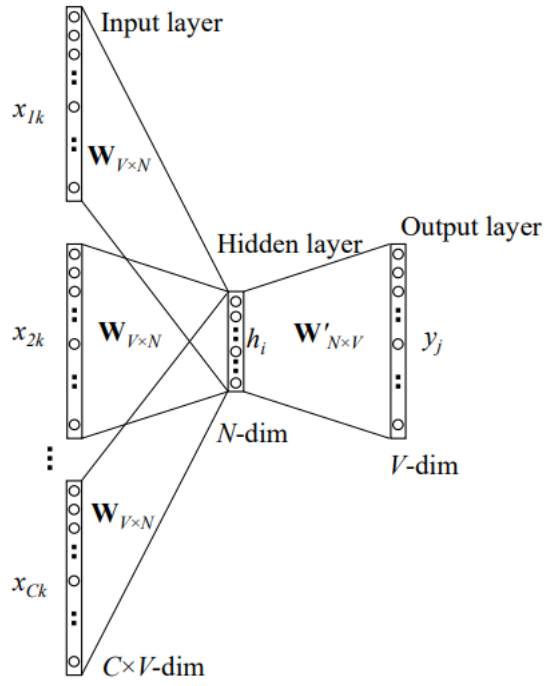


Figure 5. CBoW Model (Rong, 2014)

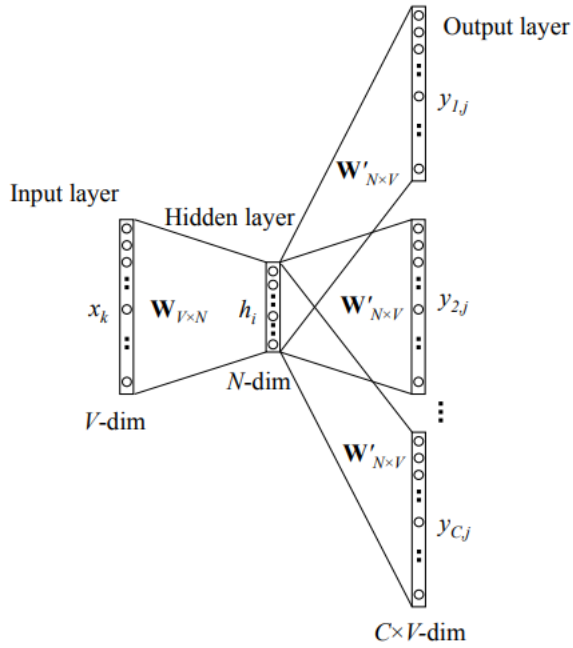


Figure 6. Skip-Gram Model (Rong, 2014)

Another neural network representation regarding Word2Vec can be seen in Figure 7. In that representation, relation/weights are tried to be learned for “women” against “queen” and “beautiful”

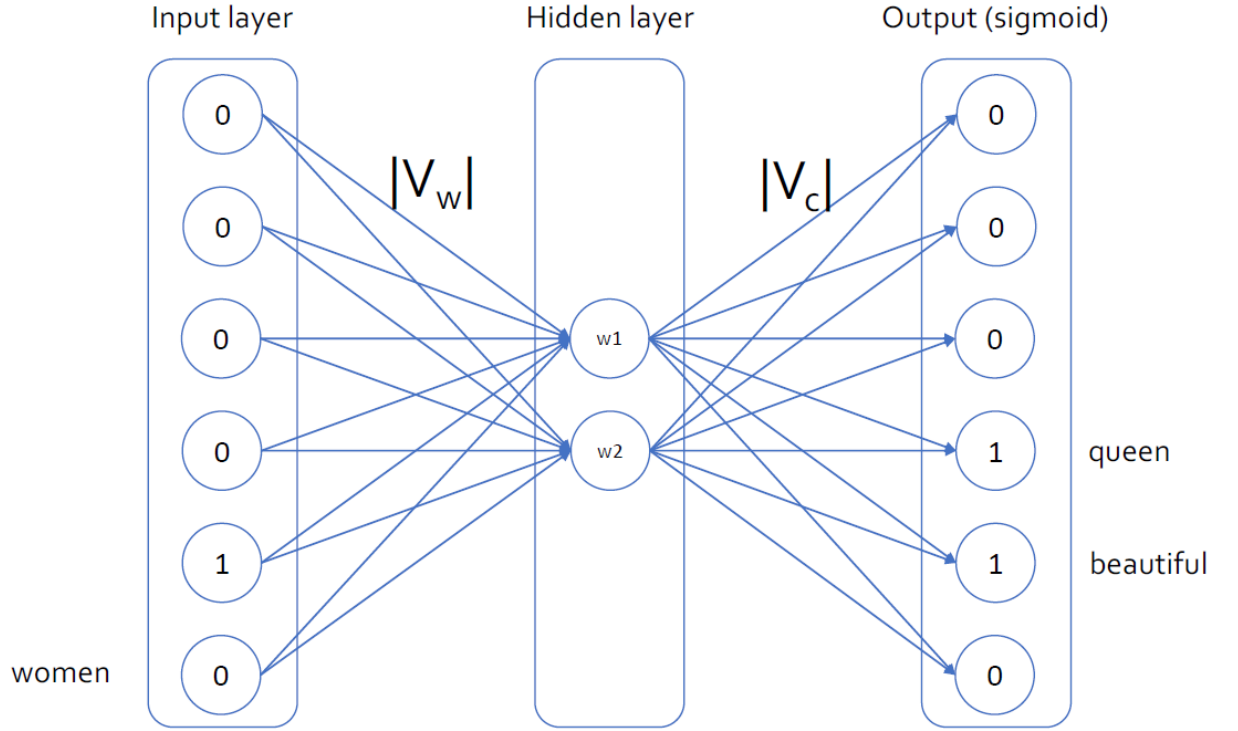


Figure 7. Sample Neural Network Representation For Word2Vec

GloVe is another word embedding model proposed by Pennington et al. (2014). The underlying approach in GloVe differs from Word2Vec by giving more importance to the frequency of co-occurrences of the words in the given text. The principal aim of GloVe is to provide word representations by utilizing the advantages of both co-occurrence matrix-based statistics and predictive models. The word vectors obtained after the training process are related to the probability of the co-occurrences of the words. A weighted least-squares model called GloVe has a log-bilinear aim. The basic premise of the model is based on the observation that ratios of word-word co-occurrence probabilities can carry some sort of meaning. Consider the co-occurrence probability of the target words stream and ice with several vocabulary probing words from a 6-billion-word corpus, here are some actual possibilities in Figure 8.

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k ice)$	$1.9 \times 10^{-4}$	$6.6 \times 10^{-5}$	$3.0 \times 10^{-3}$	$1.7 \times 10^{-5}$
$P(k steam)$	$2.2 \times 10^{-5}$	$7.8 \times 10^{-4}$	$2.2 \times 10^{-3}$	$1.8 \times 10^{-5}$
$P(k ice)/P(k steam)$	8.9	$8.5 \times 10^{-2}$	1.36	0.96

Figure 8. Actual Probabilities Table (NLP Stanford, 2002)

## 2.5 Deep Learning Algorithms

In this section, deep learning algorithms/methods used in this study are described in two sub-sections. In first sub-section, it is talked about Artificial Neural Networks (ANN). In the second sub-section, LSTM is mentioned.

**2.5.1 Artificial neural networks.** Machine learning is the process of programming computers to maximize a performance criterion based on previous experience or example data (Alpaydin, 2020). Generally, there is a model defined with some parameters. Learning phase starts with the execution of a computer program to make parameter optimization for the model utilizing the training data. The trained model can be used to make forecasts in the future. The theory of statistics is used by machine learning to build mathematical models. The main task here is generating inference from provided samples.

There are three types of machine learning problems. Unsupervised Learning aims to discover patterns in unlabeled data. In Supervised Learning, all samples are labeled. Training is performed based on labeled data. Semi-supervised Learning is located between Unsupervised Learning and Supervised Learning. During training, a little amount of labeled data is combined with a big amount of unlabeled data. In this thesis study, we generally applied Supervised Learning for both use cases.

From the point of task view, there are three types of tasks. Classification task is simply related to predicting a category of data. For example, predicting an e-mail whether if it is spam or not. Regression task is related to predict/estimate numerical values. For example, predicting stock price or house price. Clustering task deals with finding groupings of data and a label associated with each of these groupings (clusters). For example, customer segmentation. In this thesis study, for Opinion Mining use case, Classification and Clustering tasks are applied. For Opinion Forecasting, Regression and Clustering tasks are applied. A sample visualization related to these tasks can be seen in Figure 9.

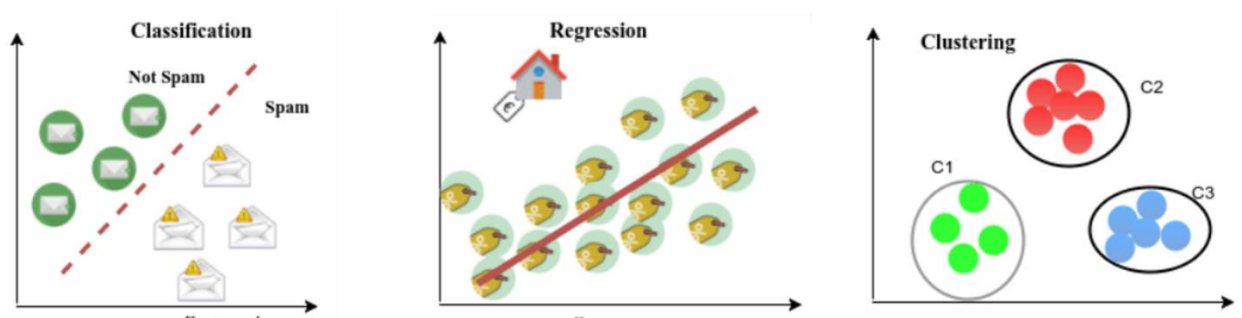


Figure 9. Machine Learning Tasks (Kayte, 2021)

ANN referred to as “neural networks” stands for making inference from the human brain working mechanism. It is inspired from the human brain. The brain consists of elements having features such as extremely complex, nonlinear, and parallel computing (information-processing system). It has the ability to organize its structural parts, known as neurons, in such a way that it can execute certain computations (such as pattern recognition, perception, and motor control) quicker than any current computer (Simon, 2009).

A biological neuron is visualized in Figure 10. It consists of synapses, dendrite, axon and cell body.

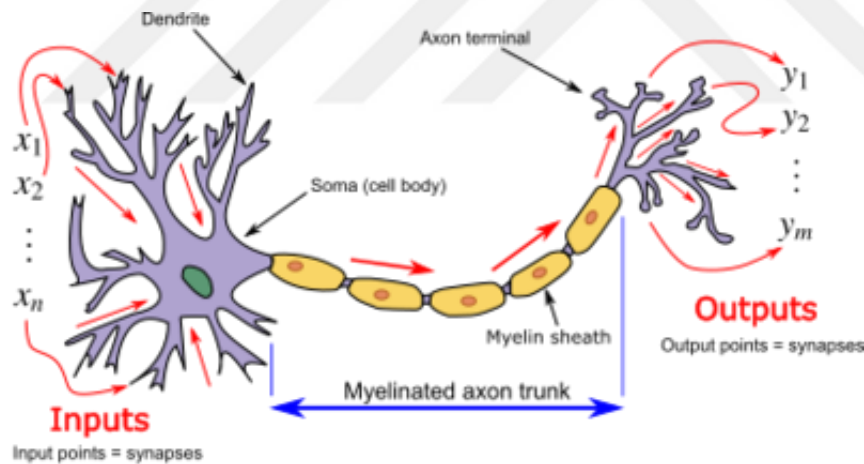


Figure 10. Biological Neuron (Wikipedia, 2022)

In Table 6, relation between Biological Neuron and Artificial Neuron is described.



Table 6  
*Relation between Biological Neuron and Artificial Neuron*

Biological Neuron	Artificial Neuron
Neuron	Processing element, node
Dendrites	Connections with propagation rule
Cell Body	Activation function, Transfer function, Output function
Axon	Connection to other neurons
Synapses	Connection weights

An artificial neuron is the fundamental unit of a DNN. An artificial neuron performs following tasks:

- takes an input,
- processes the input,
- provides it via an activation function,
- generates an output.

Warren MuCulloch and Walter Pitts proposed the first computational model of a neuron in 1943. Frank Rosenblatt, An American psychologist, introduced the traditional perceptron model in 1958. The basic processing element, the perceptron, has inputs that might originate from the environment or from the outputs of other perceptrons. Perceptron consists of 4 parts:

- Input values, an input layer
- Weights and Bias terms
- Net sum
- Activation Function (also called transfer function)

There are two types of perceptrons. Simple Perceptron is meaningful to fit linear data. For non-linear data, multilayer perceptron (MLP) structure can be used. Visualization of these perceptrons can be seen in the following figures.

In Figure 11, input  $x$  features are directly connected to target  $y$ .

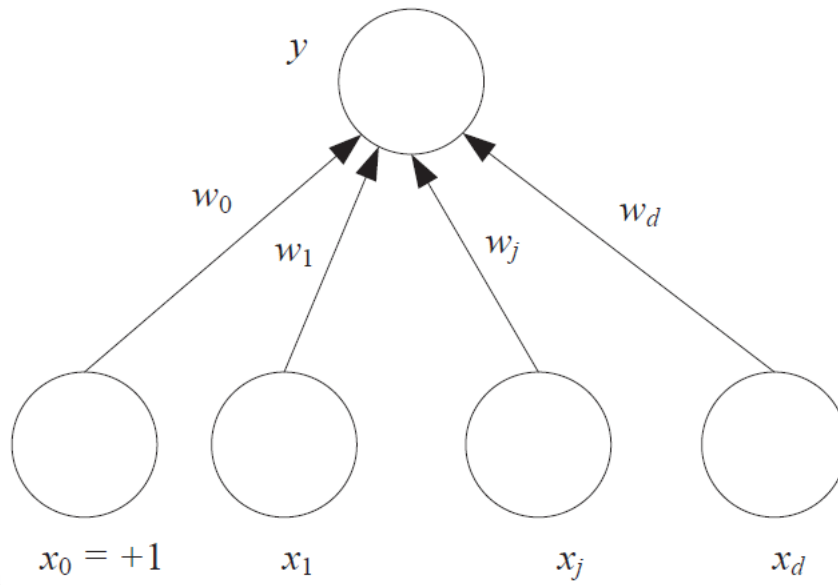


Figure 11. Simple Perceptron

In Figure 12, input  $x$  features are not directly connected to target  $y$  and there is an additional middle layer.

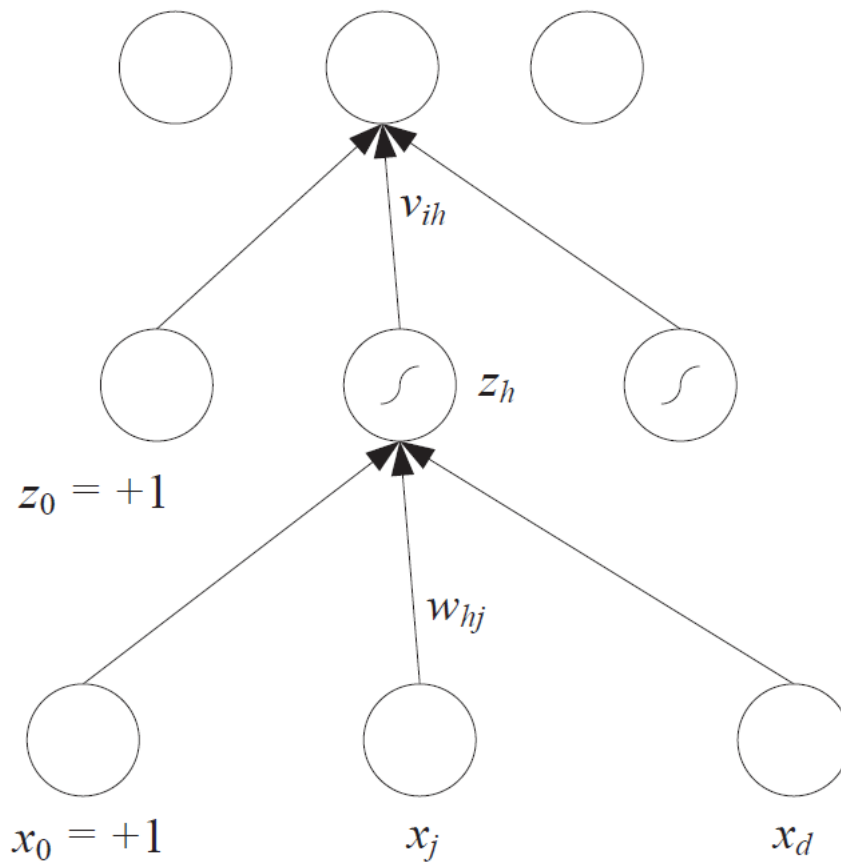


Figure 12. Multilayer Perceptron

Feed-forward networks, particularly MLPs, allow us to work with either fixed or variable length inputs, with the order of the elements ignored. When given a collection of input components, the network learns to combine them in a meaningful fashion (Goldberg, 2017).

There are several literatures works regarding utilizing of ANN method. Machine-learning techniques, such as ANN, are being used by health-care organizations to improve treatment delivery while lowering costs. Although ANN applications in diagnosis are well-known, they are increasingly being employed to influence health-care management decisions (Shahid et al., 2019). Shenfield et al. (2018) provide present a unique ANN-based method for identifying malicious network traffic that could be applied to intrusion detection systems that rely on deep packet inspection. Deep neural networks are applied to the defect detection and classification challenge to showcase their capabilities. First, the defect detection and classification tasks are formulated as classification problems using neural networks (Heo & Lee, 2018). The effects of two hyperparameters (number of hidden layers and number of neurons in the last hidden layer) as well as data augmentation on neural network performance are then examined. Escamilla-García et al. (2020) examines how ANNs are used in greenhouse technology, as well as how this sort of model might be improved in the future by incorporating new technologies such as the IoT and machine learning. Almost bulk of the works examined use a feedforward architecture, whereas recurrent and hybrid networks are underutilized in greenhouse tasks. Various network training strategies are described throughout the document, as well as the viability of applying optimization models in the learning process.

In thesis study, ANN is used in opinion mining use case for multi-class classification problem.

**2.5.2 Long short-term memory.** In the feed-forward network architecture, everything in the network looks like tied with a frozen, grade crystalline. Let us assume we would like to also allow the network's elements to change in a dynamic manner. For example, the outcome of hidden neurons may be influenced not just by past hidden layer activations, but also by earlier activations. We may also need to determine a neuron's activation phase by its own activation at aa previous time. Neural networks with that type of time-varying behavior are known as Recurrent Neural Network (RNN) (Nielsen, 2015). A simple RNN can be seen in Figure13.

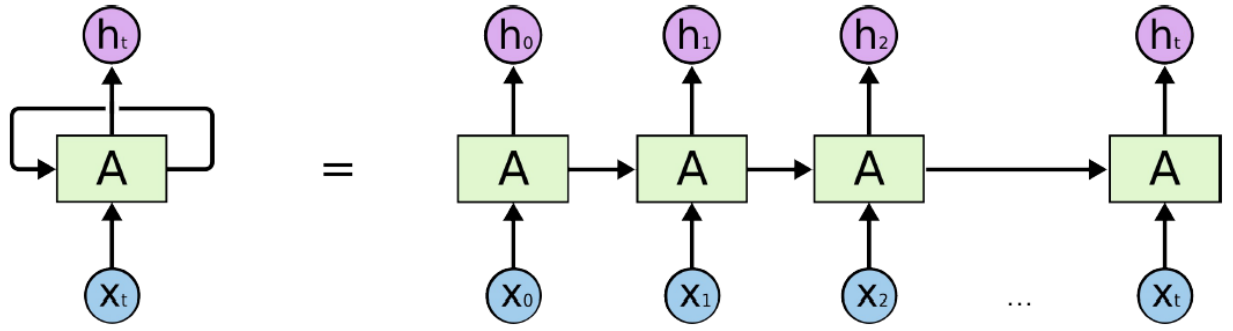


Figure 13. Simple RNN Architecture

Problems with vanishing and expanding gradients plague RNN. To solve this problem, a propose is to use the LSTM and long-term memory to alter the recurrence equation for the hidden vector. The LSTM's operations are tailored to provide fine-grained control over the data stored in this long-term memory (Aggarwal, 2018). Hochreiter and Schmidhuber proposed the LSTM cell in 1997 to address the problem of "long-term dependence". They boosted the standard recurrent cell's remembering capacity by putting a "gate" into the cell (Yu et al., 2019). Common LSTM network includes 3 gates: Forget Gate, Input Gate and Output Gate. A simple LSTM network can be seen in Figure 14.

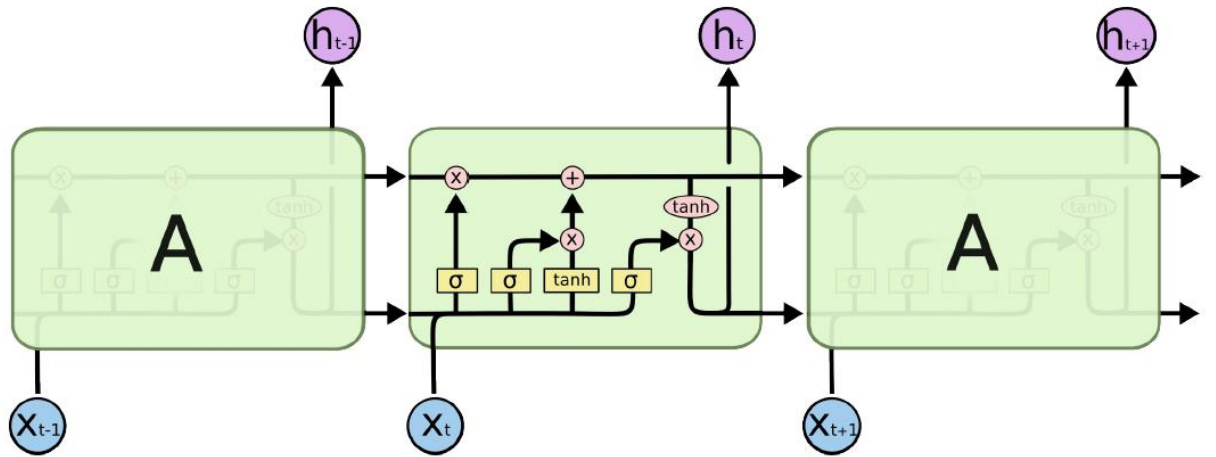


Figure 14. A Simple LSTM Architecture

There are several literatures works regarding utilizing of LSTM method. Bukhari et al. (2020) use with the support of exogenous dependent variables, a novel ARFIMA-LSTM hybrid

recurrent network captures nonlinearity in the residual values using ARFIMA model-based filters with linear tendencies better than those of the ARIMA model in the data. Proposed forecast models, which include ARIMA, support vector regression (SVR), LSTM, and bidirectional LSTM (Bi-LSTM), are evaluated in (Shahid et al., 2020) regarding the time series forecast of COVID-19 cases, deaths, and recoveries in eleven main nations. Ding et al. (2020) focus on flood forecasting in their research and offer an interpretable Spatio-Temporal Attention LSTM (STA-LSTM) based on LSTM and attention mechanisms. To develop the model, they use the dynamic attention mechanism and LSTM, the Max-Min approach to standardize data, the variable control method to pick hyperparameters, and the Adam algorithm to train it. To anticipate stock price volatility, Kim et al. (2018) presents a new hybrid LSTM model that combines the LSTM model with other generalized autoregressive conditional models.

## Chapter 3

### Models

#### 3.1 Opinion Mining for Technology Analysis with Deep Neural Networks

The primary motivation of this research is to introduce an end-to-end social media-based opinion retrieval system. In this section, we first outline the general framework of the system. Subsequently, we describe each step of the proposed framework and the dataset used in the experiments.

The general framework of the proposed system is visualized in Figure 15.

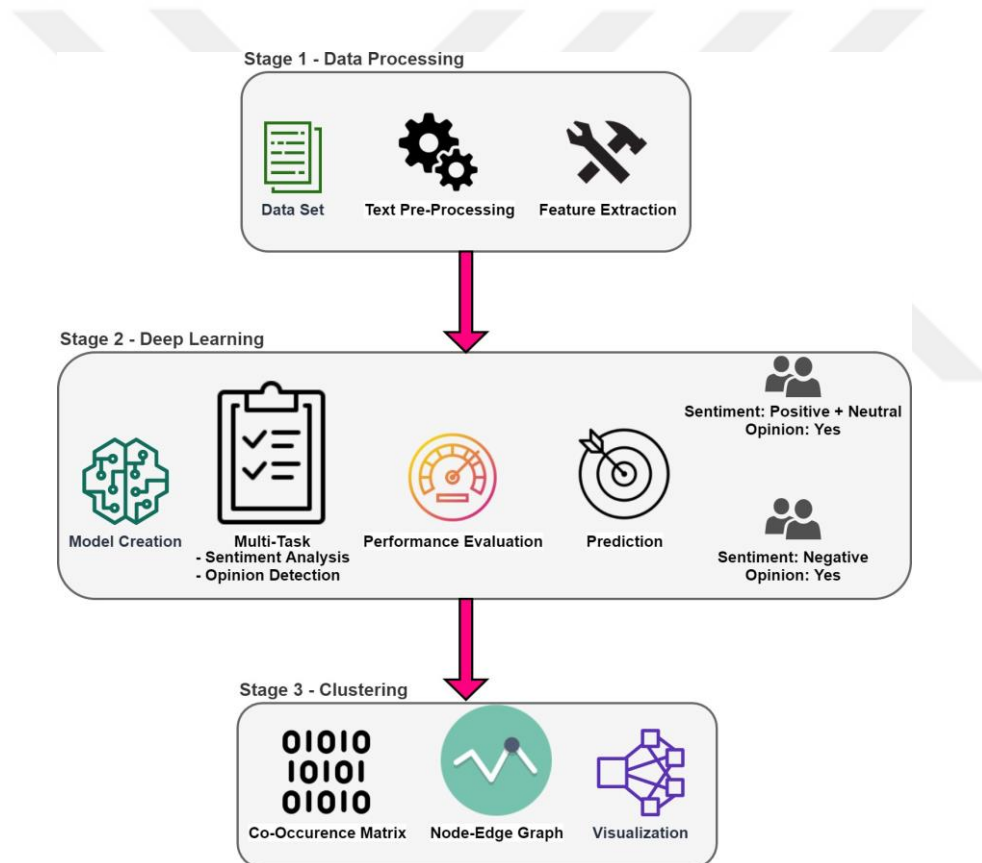


Figure 15. Methods And Operations Applied To Implement Opinion Retrieval System

During implementing the framework, Python programming language is used. Python has been around for a long time as a general-purpose programming language. Python's developer, Guido van Rossum, began working on the language in 1990. This mature and stable language is a high-level, dynamic, object-oriented language and cross-platform—all very appealing features. Python is a programming language that runs several main hardware

platforms and operating systems (Martelli et al., 2017). There are several AI and machine learning literature works utilizing Python. In this study, we used Python 3.6 version. Following libraries/packages are used/utilized within Python to be able to implement this framework.

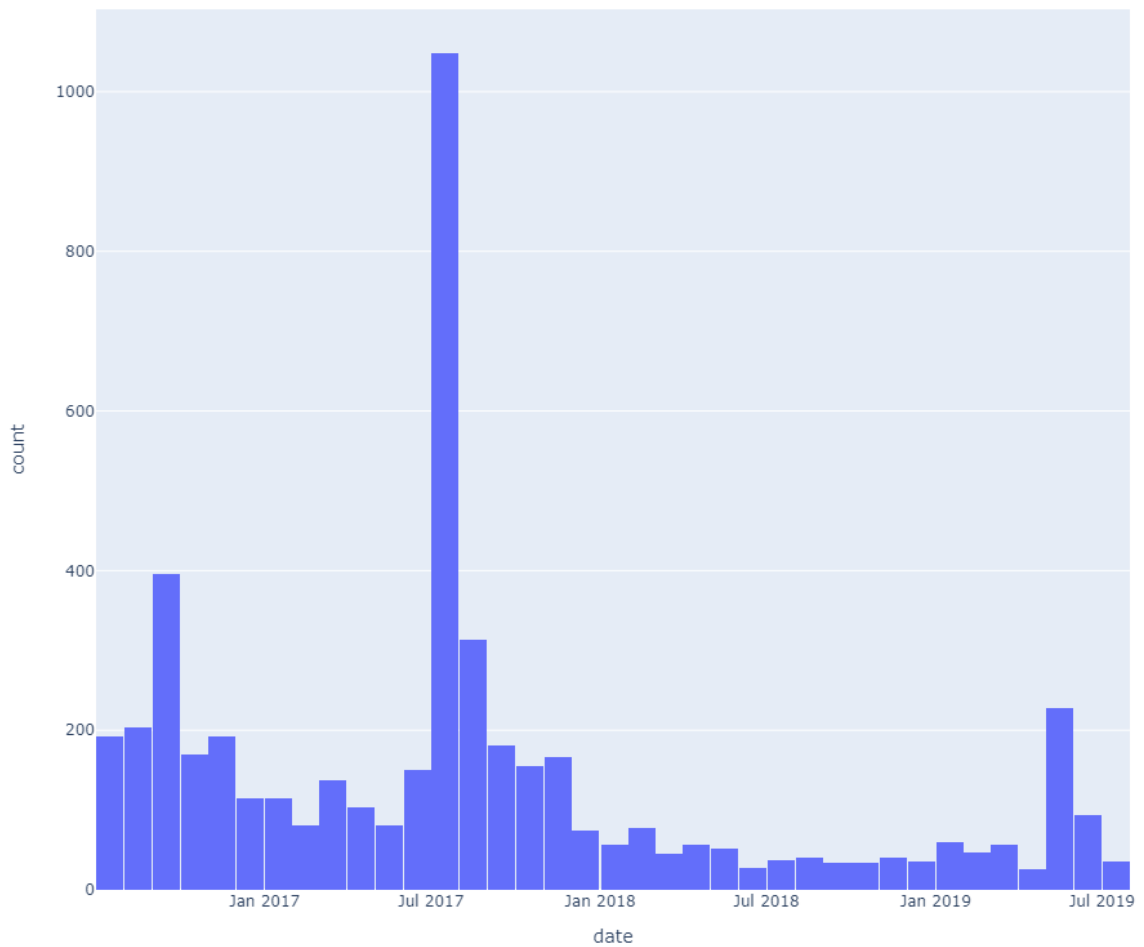
- **pandas**: It is a rapid, potent, adaptable, and user-friendly open-source data analysis and manipulation tool developed on top of the Python computer language (McKinney, 2010).
- **numpy**: One of the most important Python packages for scientific computing is this one. A multidimensional array object, derivative objects (such as masked arrays and matrices), and a number of procedures for quick array operations, including discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation, and more are all included in this Python package (Harris et al., 2020).
- **scikit-learn**: Both supervised and unsupervised learning can be supported by this machine learning library. Additionally, it features a number of tools for data preparation, model selection, and evaluation (Pedregosa et al., 2011).
- **scipy**: It is an open-source mathematics, science, and engineering software (Virtanen et al., 2020).
- **gensim**: It processes raw, unstructured digital messages using techniques from unsupervised machine learning. In order to automatically identify the semantic structure of documents, the Gensim algorithms Word2Vec, FastText, LSI, LSA, LDA, and others look at statistical co-occurrence patterns within a corpus of training documents. These algorithms are unsupervised, which means they only need a corpus of plain text documents and no human input (Rehurek & Sojka, 2010).
- **NLTK**: The most popular Python platform for interacting with data in human language is this one. It features wrappers for industrial-strength NLP libraries, an active discussion forum, a collection of text processing libraries for categorization, tokenization, stemming, tagging, parsing, and semantic reasoning, as well as simple access to over 50 corpora and lexical resources like WordNet (Bird et al., 2009).
- **matplotlib**: You can create static, animated, and interactive visualizations using this Python package. Matplotlib makes both complicated and simple tasks feasible (Hunter, 2007).

- **seaborn:** It is a Python data visualization toolkit that uses matplotlib. It provides a sophisticated interface for designing eye-catching and illuminating statistical graphics (Waskom, 2021).
- **textblob:** It is a Python package for textual data processing. It offers a basic API for doing standard NLP tasks including POS tagging, noun phrase extraction, sentiment analysis, classification, and translation, among others (Loria, 2018).
- **yellowbrick:** It improves the scikit-learn API by allowing for easier model selection and hyperparameter adjustment. Matplotlib is used behind the scenes (Bengfort et al., 2018).
- **Keras:** It is not a machine-centric API; rather, it is human-centric. Best practices for reducing cognitive load are followed by Keras, including offering consistent and straightforward APIs, reducing the number of user interactions necessary for typical use cases, and delivering clear and actionable error signals. There is a ton of documentation and developer guidelines included (Chollet & others, 2015).
- **Tensorflow:** It is a machine learning platform that is entirely open source. Its expansive, adaptable ecosystem of tools, libraries, and community resources enables developers to create and deploy machine learning applications quickly as well as academics to enhance the field's state-of-the-art (Abadi et al., 2019).
- **pickle:** For Python object structures, this module implements binary serialization and de-serialization methods. Pickling is the process of transforming a Python object hierarchy into a byte stream, and unpickling is the opposite process (Van Rossum, 2020)

**3.1.1 Dataset.** In this study, the data source was the Twitter social media platform. We collected the data using the Twitter package/tool officially published by Python Package Index (PYPI) platform. To identify the relevant data, we used a lexical search strategy, specifically, Google Glass relevant terms and hashtags. We only retrieved English tweets posted between July 2016 and July 2019. The query resulted in 4,956 unique tweets. Each tweet consists of various data fields such as username, date, year, number of retweets, text, mentions, hashtags and the original link. We used only the text content of the tweets to design a generic opinion retrieval system that can be applied to data from different social media platforms. In the following figures, some statistics regarding dataset are shared to be able to evaluate more proper.



Distribution of tweets based on date is visualized in Figure 16. According to distribution, summer period in 2017 seems having highest number of tweets.



*Figure 16. Distribution Of Tweets*

Word cloud regarding Twitter text can be seen in Figure 17. Most popular word/phrase is “Google Glass” as expected. Other major terms/words can be evaluated as “wearable tech”, “Augmented Reality”, “tech” etc.



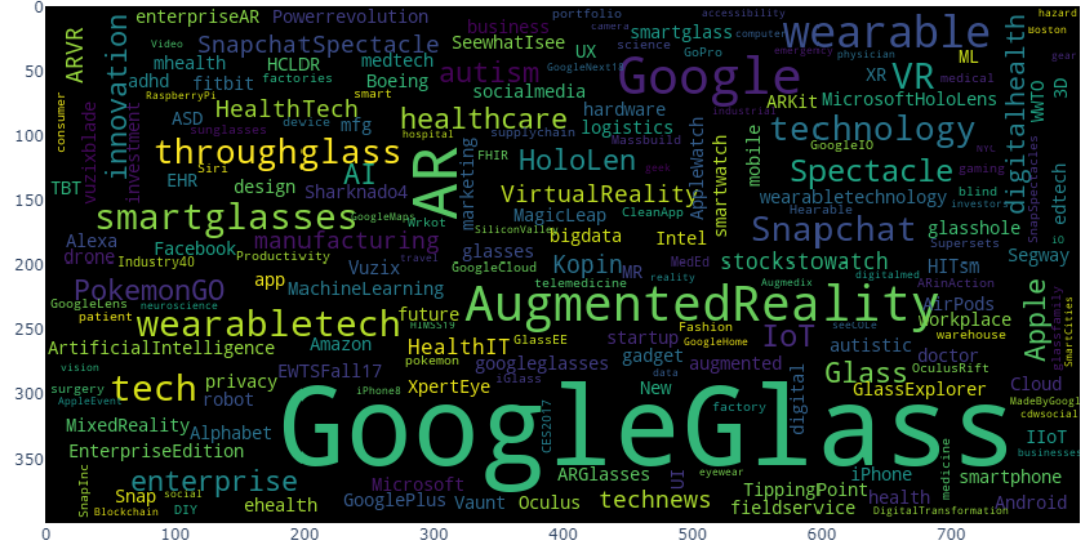


Figure 18. Word Cloud For Hashtags

After collecting the tweet dataset, we performed the manual labelling step, labelling each tweet with sentiment and opinion labels. For the labelling step, three researchers labelled 1,000 tweets separately. All labels were combined in columns side by side, and the contradicting labels were discussed and confirmed. Following the approaches listed in Table 1, we placed the sentiment target into three categories: negative, neutral, or positive whereas the opinion target is a binary variable. A tweet text entry with the label “Opinion No” indicates that the user made mention of “Google Glass” but did not give any functional feedback or suggestion. Table 7 shows the distribution of 1,000 tweets, manually labelled for the single and multi-task supervised learning problems designed to construct the opinion retrieval system. Table 8 shows some exemplary tweets along with their label information. Subsequently, we applied the DNN models built using the labelled dataset to the remaining unlabeled tweets, and both sentiment

and opinion labels were obtained for all tweets in the dataset.

Table 7  
*Class Distribution of Labelled Tweets*

Sentiment	Opinion	# Of Tweets
Negative	No	45
Negative	Yes	47
Neutral	No	165
Neutral	Yes	307
Positive	No	65
Positive	Yes	371

Table 8  
*Exemplary Labelled Tweets*

Text	Sentiment	Opinion
Looked like #googleglass failed.	Negative	No
Google Glass Dangerous for Drivers	Negative	Yes
What Happens to #Brain When You Use #GoogleGlass	Neutral	No
Imagine a future where #GoogleGlass knows how you are feeling	Neutral	Yes
Loved seeing #GoogleGlass	Positive	No
#GoogleGlass Improves #Productivity Of #Boeing #Workers	Positive	Yes

**3.1.2 Pre-processing.** For this study, we used one of the most widely used microblogging platforms, Twitter, as the data source. Twitter data have recently been used in various academic studies with sentiment analysis for various purposes. After collecting the dataset consisting of the targeted tweets, we applied several pre-processing NLP operations. First, as tweets may include the hashtag character (#), which defines the topic on the Twitter platform, we removed it to expose the real meaning of the words. We also removed the addresses of the websites, numeric forms and account names starting with characters, followed by stop words that do not contribute to the semantic meaning of the tweets. After these operations, we manually labelled some of the tweets to be used in the training of the supervised learning algorithms required for opinion retrieval task.

**3.1.3 Feature extraction.** We used Word2Vec and GloVe to obtain word vectors for

the sentiment prediction and opinion detection tasks. We used both CBoW and skip-gram implementations of Word2Vec with and without transfer learning in our experiments; the GloVe model is used only as transfer learning. In transfer learning, the main goal is to utilize a pre-trained model already built on a huge dataset for a different but similar task. In our study, the pre-trained Google News corpus word vector model consisting of 3 million 300-dimensional English word vectors and a pre-trained Twitter dataset were used with the Word2Vec and GloVe models (Rezaeinia et al., 2019).

**3.1.4 Model creation.** After representing the Twitter messages with the text representation methods, we constructed DNN-based architectures to apply the classification tasks to the extracted features. The opinion retrieval system built in this study is based on the use of two outputs for each Twitter message. The first output is sentiment label, which can be obtained by addressing a multi-class classification task including three classes which are negative, neutral and positive labels. The second output is a binary class value representing the existence of a feedback/suggestion/opinion in the related tweet.

In the model creation step, we followed two main DNN-based approaches for the abovementioned sentiment analysis and opinion detection tasks. In the first approach, we created two independent models for each of the learning tasks to obtain the sentiment and opinion outputs separately. The DNN architectures of the single task learning models are shown in Figure 19 and Figure 20. The input layer consists of the features obtained using the various word representation techniques described in Section 2.3. As is evident, each architecture has two hidden layers and two regularization techniques, drop-out and batch normalization, used after each hidden layer. In dropout, some of the randomly selected neurons are discarded in particular epochs during the learning process to enhance the network's capacity for generalization (Srivastava et al., 2014). In batch normalization, the normalization process is applied for each batch as well as in the hidden layers. The main goal is to prevent the weights and outputs from getting extreme values during the learning process (Ioffe & Szegedy, 2015).

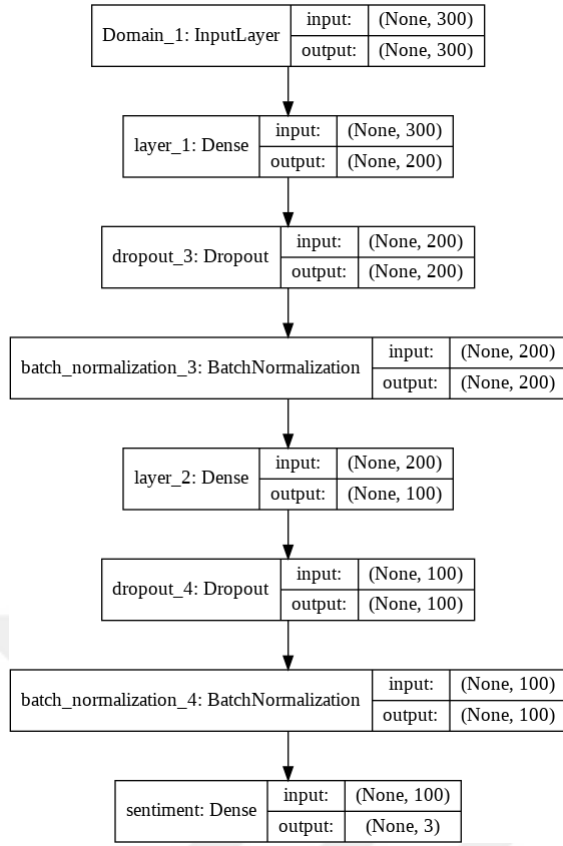


Figure 19. Single-Task DNN Architecture For Sentiment Target

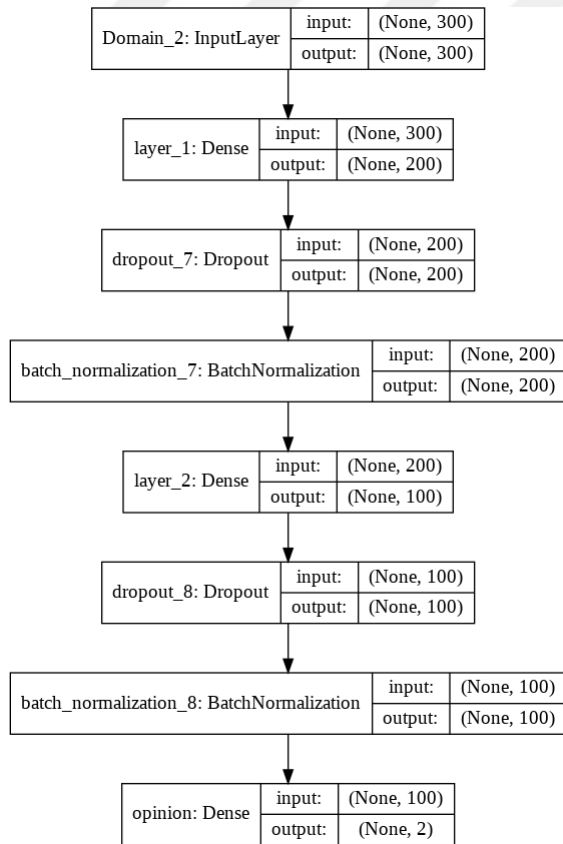


Figure 20. Single-Task DNN Architecture For Opinion Target

In the second approach, we designed a multi-task learning scheme to solve the sentiment prediction and opinion detection tasks simultaneously. Figure 21 shows one of the multi-task DNN architectures designed and tested in our study. We used a hard parameter sharing strategy in which the two NLP jobs shared access to all hidden layers. Despite the tasks having the same data source, the risk of overfitting and training time were reduced by simultaneously learning the tasks with shared layers. (Li et al., 2017; Park et al., 2019; Parwez et al., 2019). This method of learning forces the network to look for (Li et al., 2017) another regularization approach is a common representation that is capable of predicting both tasks. As seen in Figure 21, first, we concatenated features from two different text representation techniques into a single vector. We then fed the obtained representation into a dense layer which has a lower number of hidden neurons than the concatenated input layer. The dense layer was followed by dropout and normalization operations; the obtained representation was then fed into the second hidden layer which has a lower number of hidden neurons. After the second round of regularization, we mapped the representation to the specific tasks.

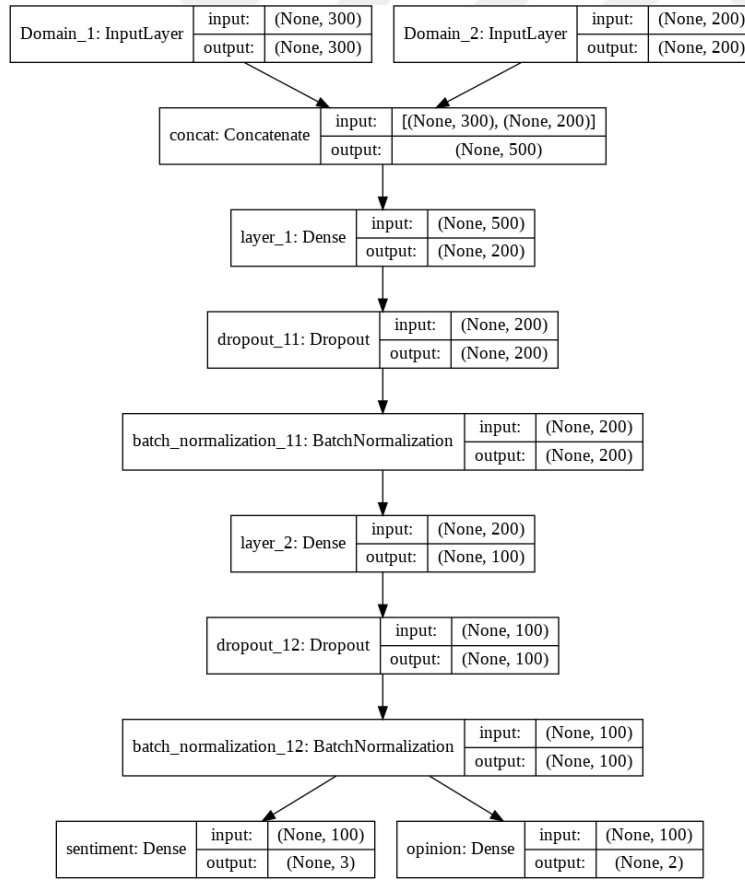


Figure 21. Multi-Task DNN Architecture Using Word2Vec + GloVe Features Together

We trained the single and multi-task learning architectures using the input space obtained with various word representation techniques. For model construction and evaluation, we initially partitioned the labelled dataset into training and test sets with 80%–20% ratios, respectively, and then divided the left-out training set into training and validation sets for hyperparameter optimization. The best model on the validation set was applied to the left-out test set. We repeated this procedure 10 times and performed statistical tests to validate the statistical significance of the comparative results. The results were evaluated in terms of F1 Score and accuracy metrics. Finally, we applied the best model to all the datasets including unlabeled tweets and obtained the sentiment and opinion predictions for use in the ultimate sentiment-based opinion retrieval task.

**3.1.5 Co-occurrence matrix-based visualization.** After obtaining the predictions for all tweets in the dataset, first, we eliminated all the tweets that did not contain any product-related feedback/suggestion/opinions as they did not contain any useful information for the opinion retrieval task. This step is a data cleaning operation which enables a better representation of opinions shared in the social media. We then used the predicted sentiment labels to divide the remaining tweets into two groups. The first group contained the tweets labelled as negative by the sentiment analysis model, representing the complaints, deficiencies, and concerns regarding the product. The second group contained the remaining positive and neutral tweets representing the customer satisfaction and expectations about the product. Accordingly, we constructed vocabulary sets for each group according to their average TF-IDF scores; a vocabulary consisting of 1-Gram and 2-Gram terms was created for this purpose. Subsequently, we created a co-occurrence matrix showing to what extent each pair of the words in the vocabulary are related for each group. Each entry of the co-occurrence matrix represents the number of times the corresponding word pair appears in the same tweet.

For the first group, we identified 70 words as vocabulary. Original co-occurrence matrix consists of 70X70 cells. In Figure 22, we aimed to display a subset. From this co-occurrence matrix, it can be said that “failure” and “learn” has a strong relation.



	failure	spectacle	time	camera	consumer	learn	game	video	driver	product
failure	0.00	0.00	0.00	0.22	0.00	0.91	0.30	0.00	0.00	0.21
spectacle	0.00	0.00	0.33	1.21	0.00	0.00	0.00	0.00	0.00	0.50
time	0.00	0.33	0.00	0.35	0.28	0.00	0.00	0.00	0.00	0.00
camera	0.22	1.21	0.35	0.00	0.00	0.00	0.00	0.50	0.00	0.00
consumer	0.00	0.00	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.28
learn	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
game	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
video	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
driver	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
product	0.21	0.50	0.00	0.00	0.28	0.00	0.00	0.00	0.00	0.00

Figure 22. Co-Occurrence Matrix For Negative Group Words

For the second group, we identified 146 words as vocabulary. Original co-occurrence matrix consists of 146X146 cells. In Figure 23, we aimed to display a subset. From this co-occurrence matrix, it can be said that “factory” and “worker”; “healthcare” and “doctor”; “manufacturing and “industry” have a strong relation.

	healthcare	autism	factory	manufacturing	doctor	ai	innovation	workplace	industry	worker
healthcare	0.00	0.09	0.00	1.70	3.00	1.20	0.81	0.00	3.10	0.29
autism	0.09	0.00	0.00	0.00	0.16	2.50	1.00	0.00	0.50	0.00
factory	0.00	0.00	0.00	2.80	0.00	1.90	1.20	0.50	0.11	3.60
manufacturing	1.70	0.00	2.80	0.00	0.00	0.67	1.90	0.00	3.70	1.40
doctor	3.00	0.16	0.00	0.00	0.00	0.11	0.31	0.00	0.00	0.00
ai	1.20	2.50	1.90	0.67	0.11	0.00	2.00	0.00	0.13	0.64
innovation	0.81	1.00	1.20	1.90	0.31	2.00	0.00	0.00	0.36	0.95
workplace	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.50	0.00
industry	3.10	0.50	0.11	3.70	0.00	0.13	0.36	0.50	0.00	0.98
worker	0.29	0.00	3.60	1.40	0.00	0.64	0.95	0.00	0.98	0.00

Figure 23. Co-Occurrence Matrix For Positive-Neutral Group Words

Finally, using each co-occurrence matrix as a complete graph, we generated and visualized the word clusters for sentiment-based opinion analysis using the force-atlas algorithm (Jacomy et al., 2014). Cosine similarity was used for clustering. For visualization, we applied a pre-processing operation known as edge-filtering (Jia et al., 2008). Related to this, a threshold value for similarity was determined and the edges above this threshold were eliminated to keep only the edges that represent strong linkage between the corresponding word pairs. We investigated the practical implications based on the obtained word clusters.

**3.1.6 Results.** To perform the experiments, we used the Google Colab platform. Google Colaboratory, often known as "Google Colab" or "Colab," is a research effort that aims to prototype using machine learning models with high-end hardware like GPUs and TPUs (Bisong, 2019). It provides an interactive Jupyter notebook environment that is serverless. Google Colab, like the rest of the G Suite, is free to use. There are several academic studies using Google Colab for verifying/testing/running developed machine learning models (Kanani & Padole, 2019; Rashid et al., 2019; Rahman et al., 2021; Gaur & Kumar, 2022; Tiwari et al., 2020). A simple screenshot from Google Colab page can be seen in the following figure. You can use this environment to run your code from scratch with using online resources of CPU, GPU and even TPU if available. You can upload and import your code and download the results.

We used Keras library with Tensorflow backend to build DNN models and Gensim library (Řehůřek & Sojka, 2010) to obtain Word2Vec and GloVe representations of the tweets. For hyper-parameter optimization, we tested various values for the hyper-parameters of feature extraction and tested model training algorithms. For this purpose, first we split the labelled tweet dataset consisting of 1,000 samples into training and test sets at 80% and 20%, respectively. Next, we used 20% of the training set as a validation set to find the optimal values of hyper-parameters for each algorithm. The best model obtained on the validation set was applied on the test set for performance evaluation. This procedure was applied 10 times and the Wilcoxon test was applied to assess the results' statistical significance.

Regarding feature extraction, we tested different values of vocabulary size from 50 to 500 to construct the feature matrices for the TF, TF-IDF, Word2Vec and GloVe methods, using both CBoW and Skip-Gram implementations of Word2Vec. For transfer learning, we also used the pre-trained word vectors of Word2Vec and GloVe. For DNN-based model creation, we tested two activation functions, tanh and ReLU, and various drop-out rates between 0.2 and 0.8. The number of hidden units in the dense layers were tested with [(600,300), (200,100),(100,50)] values. We performed the experiments for 14 different configurations of single/multi-task DNN architectures and various word representation methods.

**3.1.6.1 Experimental results.** The opinion retrieval system proposed in this study is based on two learning tasks: sentiment prediction and opinion detection. Table 9 shows the test

set performance results for the sentiment prediction task with various combinations of DNN architectures and word representation methods. As evident, the best performance, with a 0.63 F1 score and 0.66 accuracy, was obtained with a multi-task DNN architecture using Word2Vec for word representation. The second-best result, with an F1 score of 0.62 and accuracy of 0.66, was also obtained with a multi-task architecture using Word2Vec and GloVe features together. However, we should note that the difference between these two models is not significant ( $p\text{-value} > 0.05$ ). The third highest F1 score of 0.60 and accuracy of 0.65 were achieved by a single-task DNN architecture with Word2Vec features. The Wilcoxon test indicated that the difference between F1 scores of multi-task + Word2Vec and single-task + Word2Vec models is significant ( $p\text{-value} < 0.05$ ). The results also show that, in general, the DNN models based on word embedding features yield better performance compared to the models using BoW features as input.

Table 9  
*Average Performance Results on the Test Set for the Sentiment Prediction Task*

<b>Target</b>	<b>Model</b>	<b>F1 score</b>	<b>Accuracy</b>
Multi-task	Word2Vec	0.632	0.663
Multi-task	Word2Vec + GloVe	0.618	0.663
Single-task	Word2Vec	0.597	0.654
Multi-task	GloVe	0.575	0.641
Single-task	GloVe	0.558	0.646
Multi-task	TF-IDF + Word2Vec	0.507	0.605
Multi-task	TF-IDF + GloVe	0.446	0.563
Multi-task	TF + Word2Vec	0.431	0.536
Multi-task	TF + GloVe	0.413	0.529
Single-task	TF-IDF	0.376	0.475
Single-task	TF	0.370	0.470
Multi-task	TF + TF-IDF	0.366	0.466
Multi-task	TF-IDF	0.288	0.381
Multi-task	TF	0.280	0.398

The performance results regarding the opinion detection task can be seen in Table 10. Like the results obtained for the sentiment prediction task, the best two performances with an F1 score of 0.68 were yielded by the multi-task + Word2Vec and Multi-task + Word2Vec + GloVe models. The third highest F1 score was also obtained with a multi-task DNN architecture using GloVe features as input, and the performance differences among these top three models are not statistically significant ( $p\text{-value} > 0.05$ ). The best performing single-task DNN architecture for the opinion detection task yielded an F1 score of 0.65 and accuracy of

0.72 with Word2Vec features. The statistical test revealed that the F1 scores of the best performing multi-task models, multi-task + Word2Vec and Multi-task + Word2Vec + GloVe, are significantly higher than that of the best single-task model. In parallel to the results obtained for the sentiment prediction task, the word embedding features performed better than the BoW features.

Table 10

*Average Performance Results on the Test Set for the Opinion Detection Task*

<b>Target</b>	<b>Model</b>	<b>F1 score</b>	<b>Accuracy</b>
Multi-task	Word2Vec	0.677	0.719
Multi-task	Word2Vec + GloVe	0.676	0.727
Multi-task	GloVe	0.657	0.717
Single-task	Word2Vec	0.652	0.723
Single-task	GloVe	0.602	0.698
Multi-task	TF-IDF + Word2Vec	0.583	0.700
Multi-task	TF-IDF + GloVe	0.542	0.676
Multi-task	TF-IDF	0.484	0.610
Multi-task	TF + GloVe	0.474	0.661
Multi-task	TF + Word2Vec	0.472	0.668
Multi-task	TF	0.471	0.641
Single-task	TF-IDF	0.470	0.632
Multi-task	TF + TF-IDF	0.443	0.641
Single-task	TF	0.426	0.645

Regarding the results seen in Table 9 and 10, the Multi-Task + Word2Vec model yielded the highest F1 scores for both the sentiment prediction and opinion detection tasks. Therefore, we applied this model to the remaining unlabeled 3,956 tweets. The distribution of the obtained predictions with respect to the class labels are shown in Table 11.

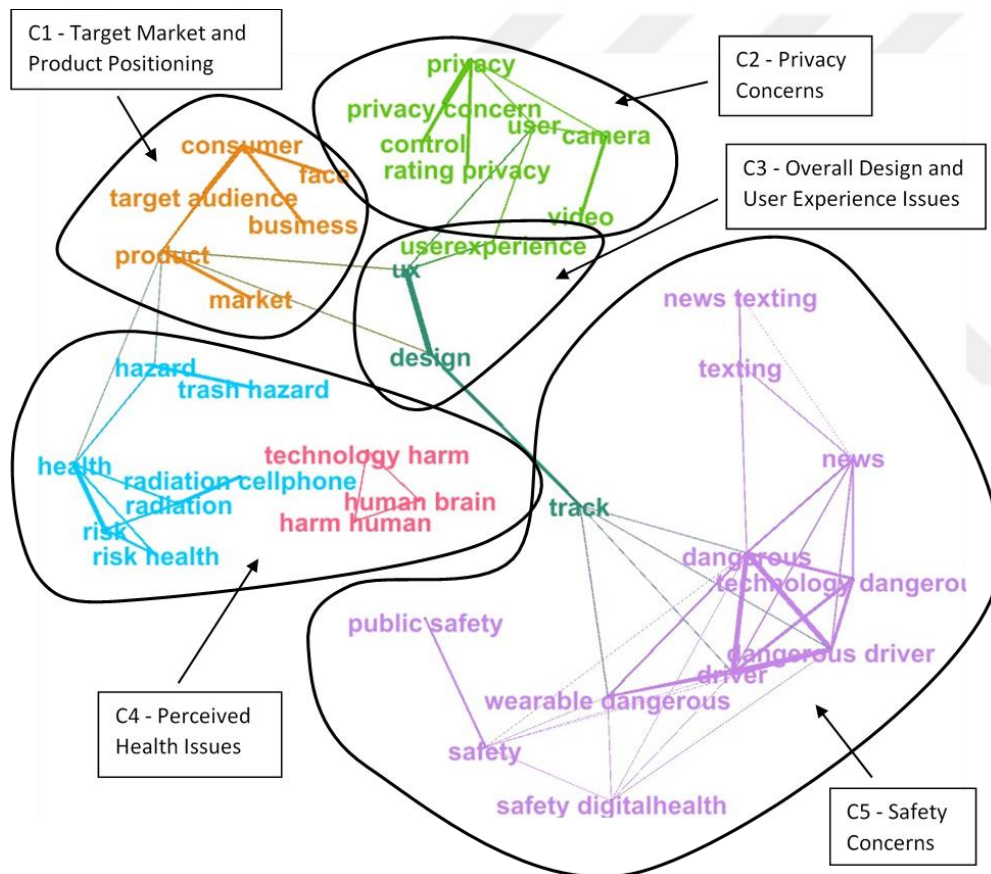
Table 11

*Class Distribution of the Tweets Labelled with the Best Multi-Task Model*

<b>Sentiment</b>	<b>Opinion</b>	<b># of Tweets</b>
Negative	No	332
Negative	Yes	227
Neutral	No	881
Neutral	Yes	1385
Positive	No	401
Positive	Yes	1730

**3.1.6.2 Practical results.** The keyword network map for each category obtained using the predictions of the best DNN model following the methodology can be seen in Figures

24 and 25. As shown in Figure 24, there are five identified major clusters which summarize negative opinions for the Google Glass. Cluster 1 (C1) shows overall problems regarding the market positioning of the Google Glass. Some consumers believe it is better to position this product for scientists and companies rather than generic consumer usage. C2 illustrates the privacy concerns for this product, as users could potentially record others without their consent. C3 reveals that many customers require a more appealing design and a better user experience. C4 illustrates some of the consumer concerns are related to the potential for brain cancer due to the wireless radiation in the device. Finally, C5 depicts the safety concerns of consumers such as the potential of the glass to limit the vision of or distract drivers, cyclists, or pedestrians in particular.



*Figure 24. Word Cluster Visualization Of The Tweets That Includes Opinion And Is Labelled As Negative By The Sentiment Prediction Model*

Figure 25 shows the clustering visualization of tweets predicted as positive or neutral by the sentiment prediction model. There are seven major clusters of the related keywords. C1

illustrates the application areas of Autism and Brain Disorder: there are various applications of this to help patients socialize and interact. C2 shows the various hardware- or software-oriented capabilities that consumers like. C3 illustrates the many application areas of Google Glass in factory conditions to mainly improve productivity and support with work tool-related applications. C4 demonstrates the application areas of this solution for medical students, such as surgical simulations. C5 groups disability-related user cases, such as supporting visually impaired individuals. C6 shows the applications of the product in surgical cases as a collaborative tool. Finally, C7 reveals the many e-health application-related suggestions and feedback to improve doctor–patient interaction.

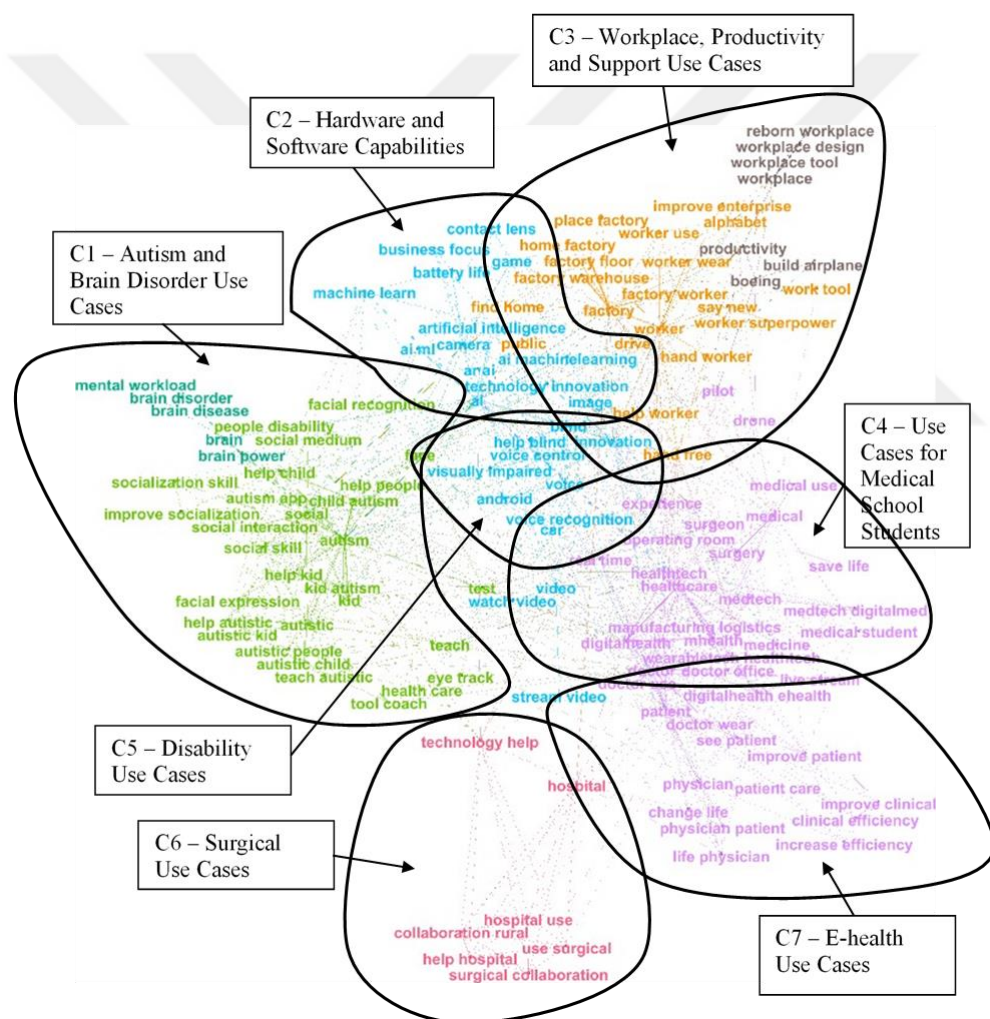


Figure 25. Word Cluster Visualization Of The Tweets That Includes Opinion Labelled As Positive Or Neutral By The Sentiment Prediction Model

Considering the results in both Figure 24 and 25, it is evident that Google Glass needs to resolve a number of issues and position itself better for commercial success. Accordingly, the

company should follow two different strategies based on different market targets and product positioning. Firstly, the next generation of Google Glass should have different versions for professionals such as for those scientists, doctors, and workplace conditions. This targeted approach can help eliminate negative feedback such as privacy, safety, and design issues. The second strategy can be for direct consumers by enhancing its hardware and software capabilities such as enhanced AI-based support, better design and the resolution of privacy and safety concerns. Examining both strategies, a niche market or targeted approach for professional use cases is a better strategy in the short or medium term for the next generations of this product. In the longer term, Google could then work towards an enhanced design for the public whilst attempting to address the public's concerns. To ensure vast technological adoption, legal issues may be the key factor for this product.

### **3.2 Technology Forecasting for Technology Analysis with Deep Neural Networks**

The primary motivation of this research is to introduce a novel approach about predicting/forecasting emerging topics regarding technology and science. In this section, we first outline the general framework of the system. Subsequently, we describe each step of the proposed framework and the dataset used in the experiments.

The general framework of the proposed system consists of the following stages:

- Stage 1: Data Collection and Processing
- Stage 2: Feature Extraction
- Stage 3: Forecasting with Deep Neural Networks
- Stage 4: Visualization and Quantitative Analysis

The methods and operations applied to implement each step of the general framework are shown in Figure 26.

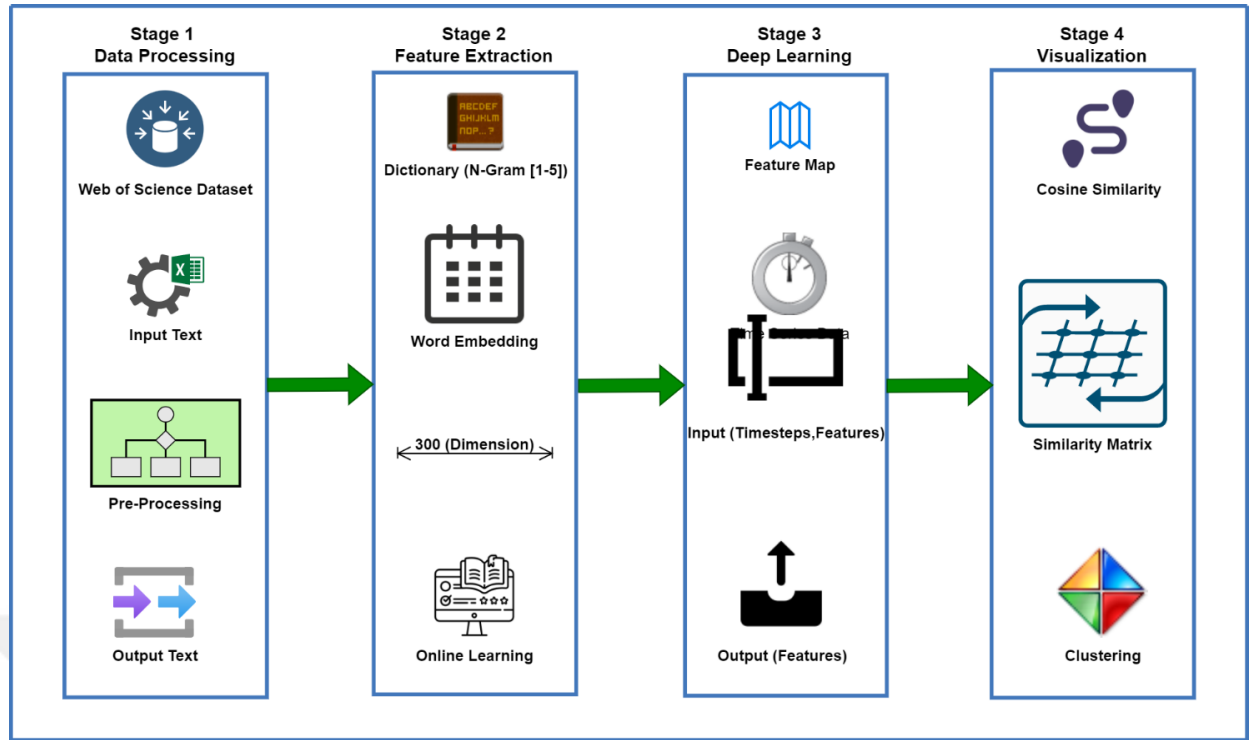


Figure 26. General Framework

In addition to Python libraries aforementioned in 3.1, following Python libraries are also utilized.

- **pandasgui**: It is a graphical user interface for looking at, charting, and analyzing Pandas Data Frames.
- **selenium.webdriver**: A straightforward API is offered by the Selenium Python bindings for creating Selenium WebDriver functional and acceptability tests. You may easily and naturally access all of Selenium WebDriver's functionality by using the Selenium Python API.
- **pyLDavis**: It's been created to help people understand the topics in a topic model that has been fitted to a corpus of text data. The software builds an interactive web-based representation using information from a fitted LDA topic model (Sievert & Shirley, 2014).

**3.2.1 Dataset.** Web of Science (WoS) database is used as source of dataset. It is a paid-access platform that allows users to access several databases that contain reference and citation data from academic journals, conference proceedings, and other publications in a variety of academic subjects (usually over the internet). It was created by the Institute for Scientific Information in the first place. Clarivate owns it now (previously the Intellectual Property and



Science business of Thomson Reuters). You can prepare several query combinations and filter your results.

In the beginning phase for this part, “Biochemistry & Molecular Biology” subject is selected as an interdisciplinary research area. Prepared a dataset with downloading all the articles, papers which are indexed under WoS database through 1975 – 2020. There were totally 2M+ records as number of articles/papers. Most of the articles indexed before 1991 does not have content for “Abstract”. Regarding to this, 1991 – 2020 year range is taken into account including 1.5M records. Initially, the aim was creating a co-occurrence matrix based on term frequency per month basis. However, dimension is very high (4096 X 4096) for that case. This brings out some issues related to computing power and memory processing. To overcome this bottleneck, we decided to create features based on word embedding methods.

In later phase, Text Mining sub-domain under Computer Science, Artificial Intelligence is selected as use case. As source of dataset, WoS database is used to search articles/papers to be collected. For identifying documents in most promising and correct way, search keywords/phrases are investigated and analyzed with the help of several studies in literature (Berry & Castellanos, 2007; Aggarwal, 2015; Talib et al., 2016; Weiss et al., 2010; Dang & Ahmad, 2015; Chandrayan & Bamne, 2021; Thakur & Kumar, 2020; Khan et al., 2020; Hassani et al., 2020; Kumar et al., 2021; Jung & Lee, 2020). There are numerous text mining approaches that can be used to investigate text patterns and the mining process (Gupta & Lehal, 2009). Various text mining techniques and core functionalities are visualized in Figure 27. Central points in the circles are “Information Retrieval”, “Document Clustering”, “Concept Extraction”, “Web Mining”, “Document Classification”, “Information Extraction”, “Natural Language Processing”.



	OR "Parts-of-speech" OR "Bag of Words" OR "N-Gram" OR "Syntactic Analysis" OR "Semantic Analysis") AND ("Machine Learning" OR "Data Science" OR "Data Mining" OR "Deep Learning" OR "Pattern Recognition"))	
6	TS=("Natural Language Processing") OR TS=( ("Sentiment Analysis" OR "Text Mining" OR "Text Intelligence" OR "Knowledge Retrieval" OR "Information Retrieval" OR "Tokenization" OR "Normalization" OR "Stemming" OR "Lemmatization" OR "Corpus" OR "Stop Words" OR "Parts-of-speech" OR "Bag of Words" OR "N-Gram" OR "Syntactic Analysis" OR "Semantic Analysis") AND ("Machine Learning" OR "Data Science" OR "Data Mining" OR "Deep Learning" OR "Pattern Recognition" OR "Artificial Intelligence"))	38,188
7	TS=("Natural Language Processing") OR TS=( ("Sentiment Analysis" OR "Text Mining" OR "Text Intelligence" OR "Knowledge Retrieval" OR "Information Retrieval" OR "Tokenization" OR "Normalization" OR "Stemming" OR "Lemmatization" OR "Corpus" OR "Stop Words" OR "Parts-of-speech" OR "Bag of Words" OR "N-Gram" OR "Syntactic Analysis" OR "Semantic Analysis") AND ("Machine Learning" OR "Data Science" OR "Data Mining" OR "Deep Learning" OR "Pattern Recognition" OR "Artificial Intelligence" OR "Data Analytics"))	38,496
8	TS=("Natural Language Processing") OR TS=( ("Sentiment Analysis" OR "Text Mining" OR "Text Intelligence" OR "Knowledge Retrieval" OR "Information Retrieval" OR "Tokenization" OR "Normalization" OR "Stemming" OR "Lemmatization" OR "Corpus" OR "Stop Words" OR "Parts-of-speech" OR "Bag of Words" OR "N-Gram" OR "Syntactic Analysis" OR "Semantic Analysis") AND ("Machine Learning" OR "Data Science" OR "Data Mining" OR "Deep Learning" OR "Pattern Recognition" OR "Artificial Intelligence" OR "Data Analytics" OR "Neural Network"))	42,670
9	TS=("Natural Language Processing" OR "Sentiment Analysis")  OR  (ALL=("Natural Language Processing" OR "Sentiment Analysis")  AND TS=("Sentiment Analysis" OR "Text Mining" OR "Text Intelligence" OR "Knowledge Retrieval" OR "Information Retrieval" OR "Tokenization" OR "Normalization" OR "Stemming" OR "Lemmatization" OR "Corpus" OR "Stop Words" OR "Parts-of-speech" OR "Bag of Words" OR "N-Gram" OR "Syntactic Analysis" OR "Semantic Analysis")  AND TS= ("Machine Learning" OR "Data Science" OR "Data Mining" OR "Deep Learning" OR "Pattern Recognition" OR "Artificial Intelligence" OR "Data Analytics" OR "Neural Network"))	32,037
10	TS=("Natural Language Processing") OR  (ALL=("Natural Language Processing")  AND TS=("Sentiment Analysis" OR "Text Mining" OR "Text Intelligence" OR "Knowledge Retrieval" OR "Information Retrieval" OR "Tokenization" OR "Normalization" OR "Stemming" OR "Lemmatization" OR "Corpus" OR "Stop Words" OR "Parts-of-speech" OR "Bag of Words" OR "N-Gram" OR "Syntactic Analysis" OR "Semantic Analysis")  AND TS= ("Machine Learning" OR "Data Science" OR "Data Mining" OR "Deep Learning" OR "Pattern Recognition" OR "Artificial Intelligence" OR "Data Analytics" OR "Neural Network"))	22,738
11	TS= ( "Text Mining" OR "Text Analysis" OR "Data Mining" OR "Web Mining" OR "Machine Learning" OR "Pattern Recognition" OR "Statistical Learning" OR "Deep Learning" OR "Topic Modeling" )  AND  ALL= ( "Text Summarization" OR "Text Categorization" OR	15,713

	"Text Extraction" OR "Text Clustering" OR "Text Visualization" OR "Text Classification" OR "Document Classification" OR "Document Retrieval" OR "Document Clustering" OR "Information Extraction" OR "Natural Language Processing" )	
12	TS= ( "Text Mining" OR "Text Analysis" OR "Text Analytics" OR "Text Intelligence" OR "Data Mining" OR "Data Analysis" OR "Data Analytics" OR "Web Mining" OR "Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Pattern Mining" OR "Statistical Learning" OR "Topic Modeling" OR "Information Retrieval" )  AND  ALL= ( "Text Summarization" OR "Text Classification" OR "Text Extraction" OR "Text Clustering" OR "Text Visualization" OR "Text Categorization" OR "Document Retrieval" OR "Document Clustering" OR "Document Classification" OR "Information Extraction" )	35,117

	OR "Knowledge Retrieval" OR "Knowledge Discovery" OR "Syntactic Analysis" OR "Semantic Analysis" OR "Sentiment Analysis" OR "Natural Language Processing" )	
13	TS= ( "Text Mining" OR "Text Analysis" OR "Text Analytics" OR "Text Intelligence" OR "Text Summarization" OR "Text Classification" OR "Text Extraction" OR "Text Clustering" OR "Text Visualization" OR "Text Categorization" OR "Data Mining" OR "Data Analytics" OR "Web Mining" OR "Pattern Mining" OR "Document Retrieval" OR "Document Clustering" OR "Document Classification" OR "Information Retrieval" OR "Information Extraction" OR "Knowledge Retrieval" OR "Knowledge Discovery" )  AND  ALL= ( "Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Pattern Recognition" OR "Statistical Learning" OR "Syntactic Analysis" OR	45,924

	"Semantic Analysis" OR "Sentiment Analysis" OR "Natural Language Processing" )	
14	(TS= ( "Text Mining" OR "Text Analysis" OR "Text Analytics" OR "Text Intelligence" OR "Text Summarization" OR "Text Classification" OR "Text Extraction" OR "Text Clustering" OR "Text Visualization" OR "Text Categorization" OR "Web Mining" OR "Document Retrieval" OR "Document Clustering" OR "Document Classification" OR "Document Ranking" OR "Information Retrieval" OR "Information Extraction" OR "Knowledge Retrieval" OR "Knowledge Discovery" OR "Concept Extraction" OR "Natural Language Processing" OR "Sentiment Analysis" OR "Topic Modeling" )  AND  ALL= ( "Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Pattern Recognition" OR "Neural Network" OR "Statistical Learning" OR "Syntactic Analysis" OR "Semantic Analysis" )	48,603

	OR "Sentiment Analysis" OR "Natural Language Processing" OR "Web Crawling" OR "Tokenization" OR "Stemming" OR "Lemmatization" OR "Stop Words" OR "Parts-of-speech" OR "Bag of Words" OR "N-Gram" ) )  NOT TI=("Review" OR "Survey" OR "State-of-The-Art" OR "State of The Art")  NOT TS=("Image Processing" OR "Computer Vision" OR "Object Recognition" OR "Image Classification" OR "Face Recognition")	
15	(TS= ( "Text Mining" OR "Text Analysis" OR "Text Analytics" OR "Text Intelligence" OR "Text Summarization" OR "Text Classification" OR "Text Extraction" OR "Text Clustering" OR "Text Visualization" OR "Text Categorization" OR "Text Retrieval" OR "Web Mining" OR "Document Retrieval" OR "Document Clustering" OR "Document Classification" OR "Document Ranking" OR "Document Categorization" OR "Document Summarization" OR "Information Retrieval" OR "Information Extraction" OR "Knowledge Retrieval" OR "Knowledge Discovery" OR "Concept Extraction"	50,253

OR "Natural Language Processing" OR "Sentiment Analysis" OR "Topic Modeling" OR "Topic Tracking" OR "Topic Detection" )  AND  ALL= ( "Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Pattern Recognition" OR "Neural Network" OR "Statistical Learning" OR "Syntactic Analysis" OR "Semantic Analysis" OR "Sentiment Analysis" OR "Natural Language Processing" OR "Web Crawling" OR "Tokenization" OR "Stemming" OR "Lemmatization" OR "Stop Words" OR "Parts-of-Speech" OR "Bag of Words" OR "TF-IDF" OR "Word Embedding" OR "N-Gram" OR "Phrase Recognition" OR "Entity Extraction" OR "Named Entity Recognition" ) )  NOT TI=("Review" OR "Survey" OR "State-of-The-Art" OR "State of The Art")  NOT TS=("Image Processing" OR "Computer Vision" OR "Object Recognition" OR "Image Classification" OR "Face Recognition" OR "Object Detection" OR "Edge Detection" OR "Machine Vision" OR "Image Segmentation")	
--	--

Based on analysis of literature study, following query is composed/constructed to be used



on WoS database. From the point of WoS query terminology, TS stands for searching for topic term in the fields (Title, Abstract, Author Keywords, Keywords Plus) within a record. TI stands for searching the Title field within a record.

(  
*TS=("Text Mining" OR "Text Analysis" OR "Text Analytics" OR "Text Intelligence" OR "Text Summarization" OR "Text Classification" OR "Text Extraction" OR "Text Clustering" OR "Text Visualization" OR "Text Categorization" OR "Text Retrieval" OR "Web Mining" OR "Document Retrieval" OR "Document Clustering" OR "Document Classification" OR "Document Ranking" OR "Document Categorization" OR "Document Summarization" OR "Information Retrieval" OR "Information Extraction" OR "Knowledge Retrieval" OR "Knowledge Discovery" OR "Concept Extraction" OR "Natural Language Processing" OR "Sentiment Analysis" OR "Topic Modeling" OR "Topic Tracking" OR "Topic Detection")*

*AND*

*ALL=("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Pattern Recognition" OR "Neural Network" OR "Statistical Learning" OR "Syntactic Analysis" OR "Semantic Analysis" OR "Sentiment Analysis" OR "Natural Language Processing" OR "Web Crawling" OR "Tokenization" OR "Stemming" OR "Lemmatization" OR "Stop Words" OR "Parts-of-Speech" OR "Bag of Words" OR "TF-IDF" OR "Word Embedding" OR "N-Gram" OR "Phrase Recognition" OR "Entity Extraction" OR "Named Entity Recognition"*

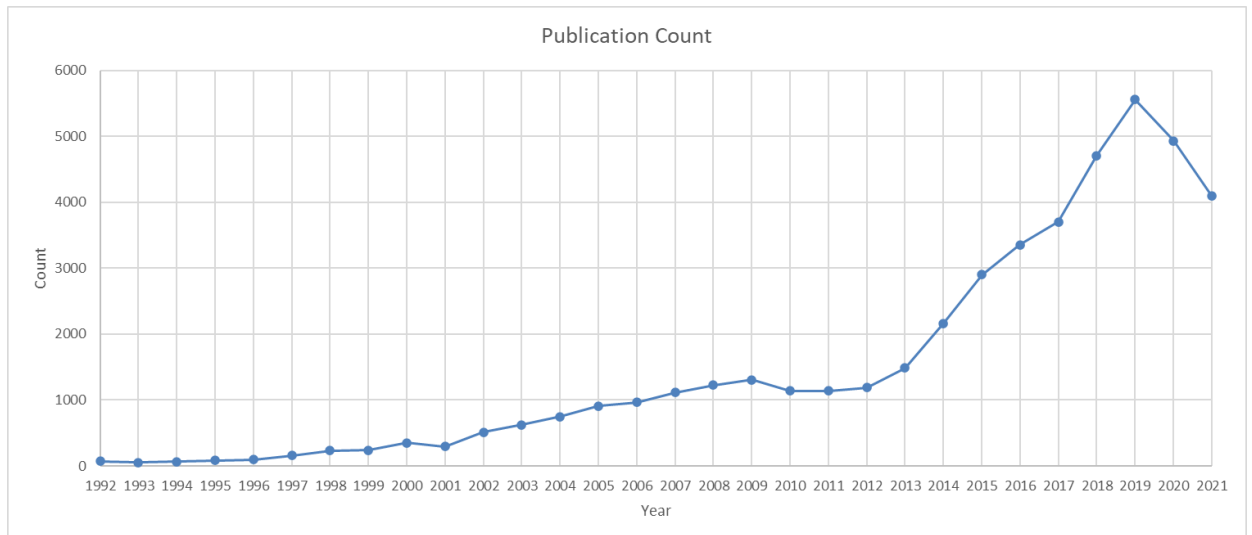
*)*

*)*

*NOT TI=("Review" OR "Survey" OR "State-of-The-Art" OR "State of The Art")*

*NOT TS=("Image Processing" OR "Computer Vision" OR "Object Recognition" OR "Image Classification" OR "Face Recognition" OR "Object Detection" OR "Edge Detection" OR "Machine Vision" OR "Image Segmentation")*

According to determined query, approximately 50.000 articles/papers are found between early 1980s and 2021. As we would like to perform online training in monthly fashion, we used this dataset as starting from 1992 January. Regarding this, utilized publication count is calculated as 45406. Distribution of publication count against years can be seen in Figure 28.

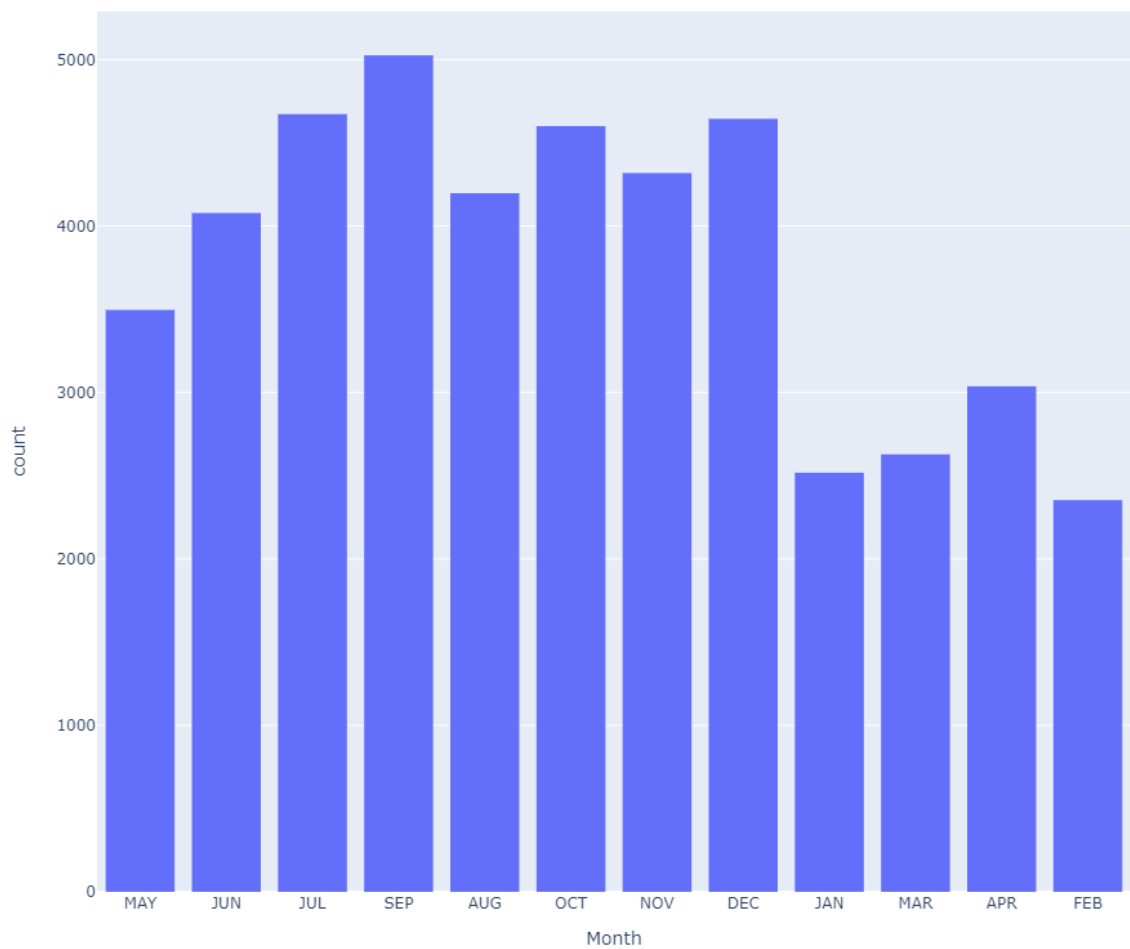


*Figure 28. Distribution Of Publication Count*

For automating download operation of found documents, a Python application is implemented with the support of Selenium package (Gundecha, 2014). Selenium is an open-source umbrella project that offers a selection of libraries and tools for automating browsers. Without needing to learn a test scripting language, it provides you with a replay tool for building functional tests. We used WebDriver version of Selenium to combine and manage via Python script. WebDriver drives a browser natively, much like a user would, either locally or remotely using the Selenium server, which is a significant step forward in browser automation. Selenium WebDriver refers to the language bindings and implementations of the specific browser-controlling code. WebDriver is the most frequent name for it.

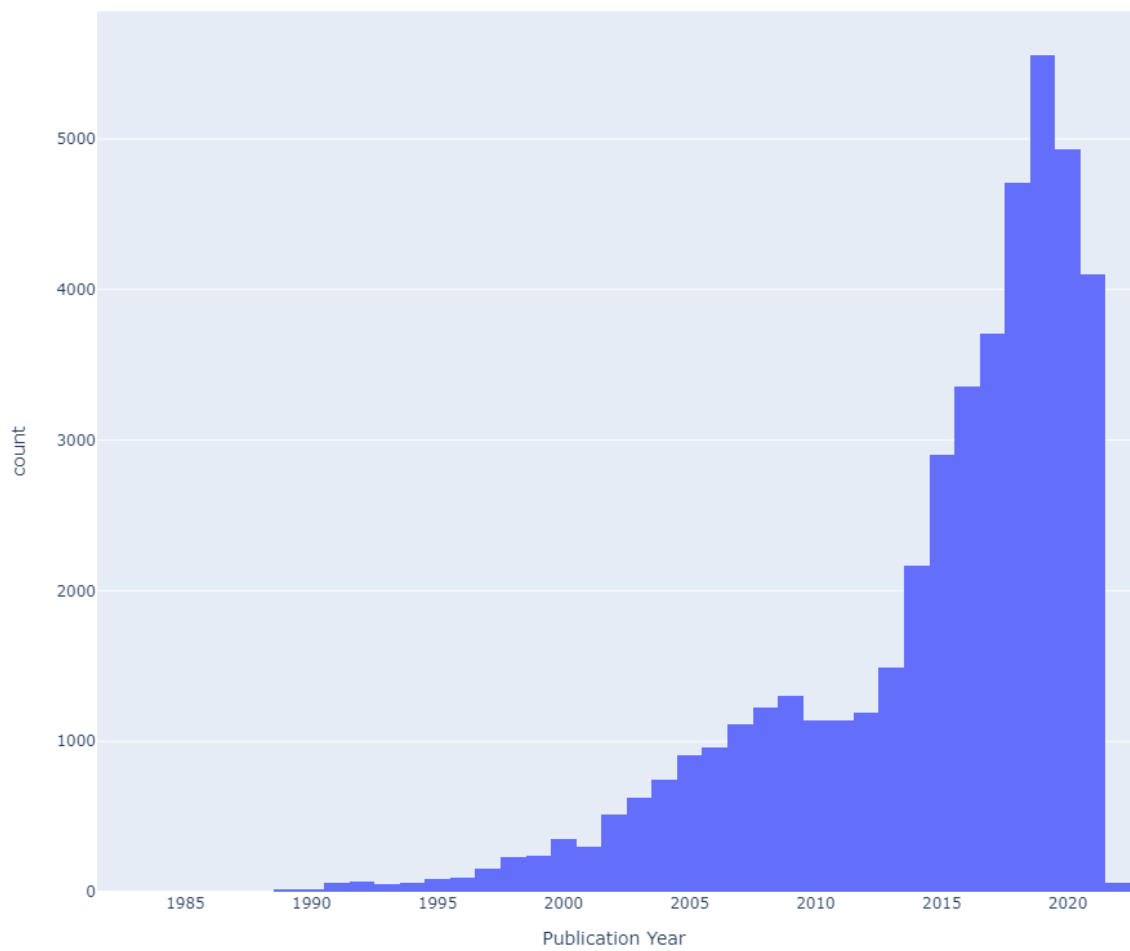
In the following figures, some statistics regarding dataset are shared to be able to evaluate data in a proper way.

Distribution of articles based on month is visualized in Figure 29. Most of the articles seem published/accepted in September. Least of the articles seem published/accepted in January and February.



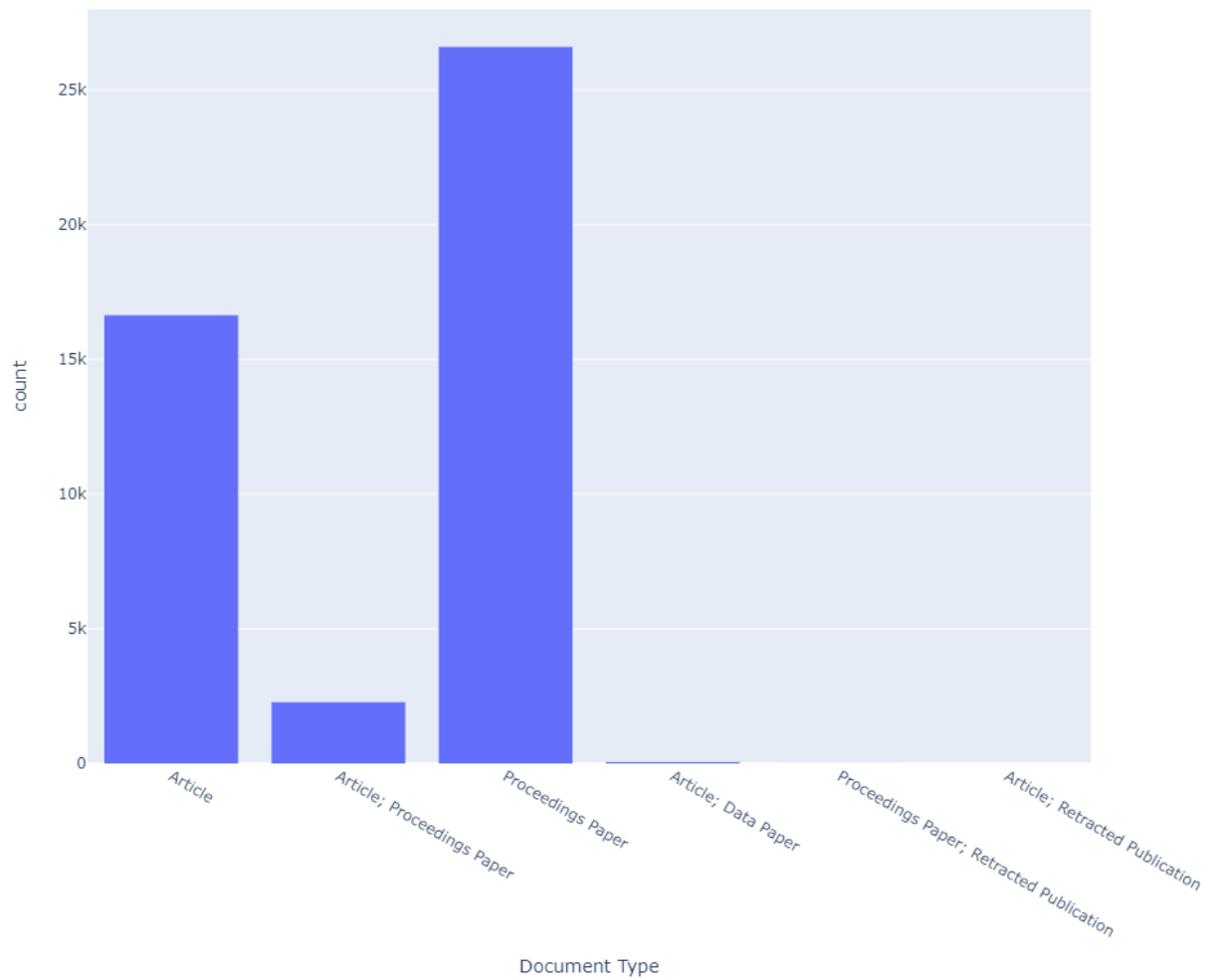
*Figure 29. Number Of Articles Per Month*

Distribution of articles based on year is visualized in Figure 30. Most of the articles seem published/accepted in 2019. Least of the articles seem published/accepted in the middle 1990s as expected.



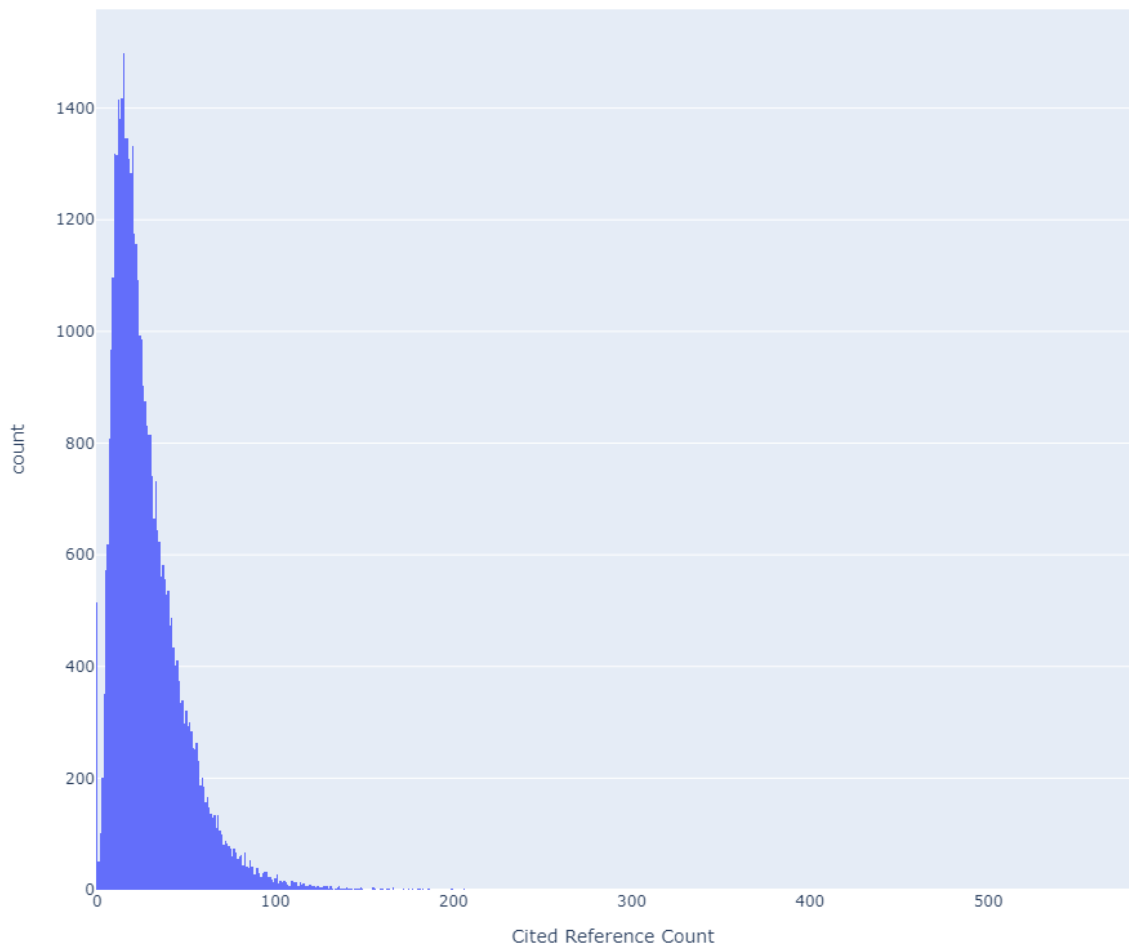
*Figure 30.* Number Of Articles Per Year

Distribution of document type is visualized in Figure 31. Most of the documents belong to Proceedings Paper and Article categories as expected.



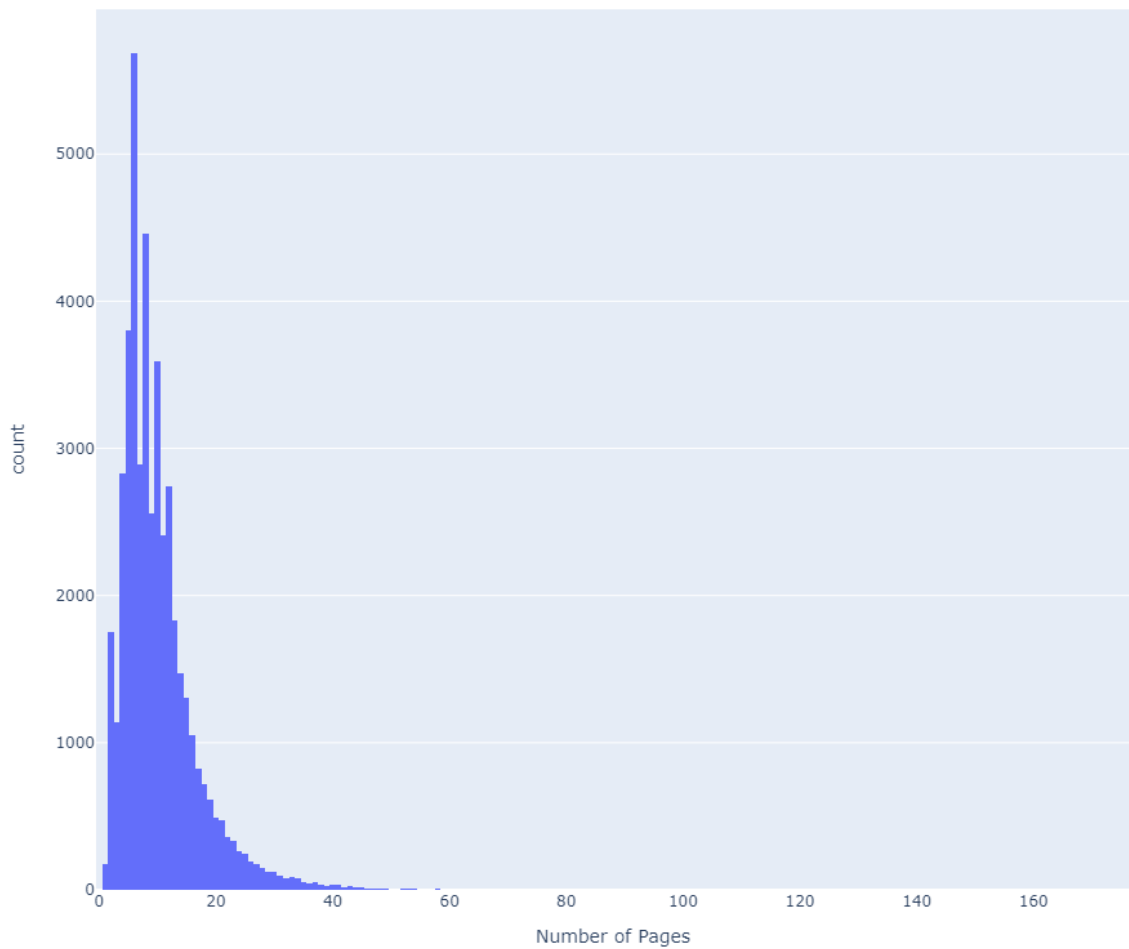
*Figure 31. Number Of Documents Per Type*

Distribution of cited reference count is visualized in Figure 32. Dominant cited reference count seems 15. Average cited reference count is 28. That tells us that each article in the dataset has been utilized in the literature system by the other academic works. In addition to, maximum reference count seems around 150 – 200.



*Figure 32. Cited Reference Count*

Distribution of number of pages count is visualized in Figure 33. Most of the articles seems having 6 pages. In addition to, average page count is 10. This results that each publication in the dataset has a good amount of work from the point of page size. As most of the documents in Proceedings Paper type, having this page number seems reasonable.



*Figure 33.* Distribution Of Number Of Pages

Word cloud regarding ‘Article Title’ is visualized in Figure 34. Dominant words/phrases are seen as “Analysis”, “Text”, “Learning”, “Classification”, “Document”, “Data”. Regarding selected domain/use-case “Text Mining”, having these dominant words seem reasonable. In addition to, having “Classification” also tells us that major machine learning task.

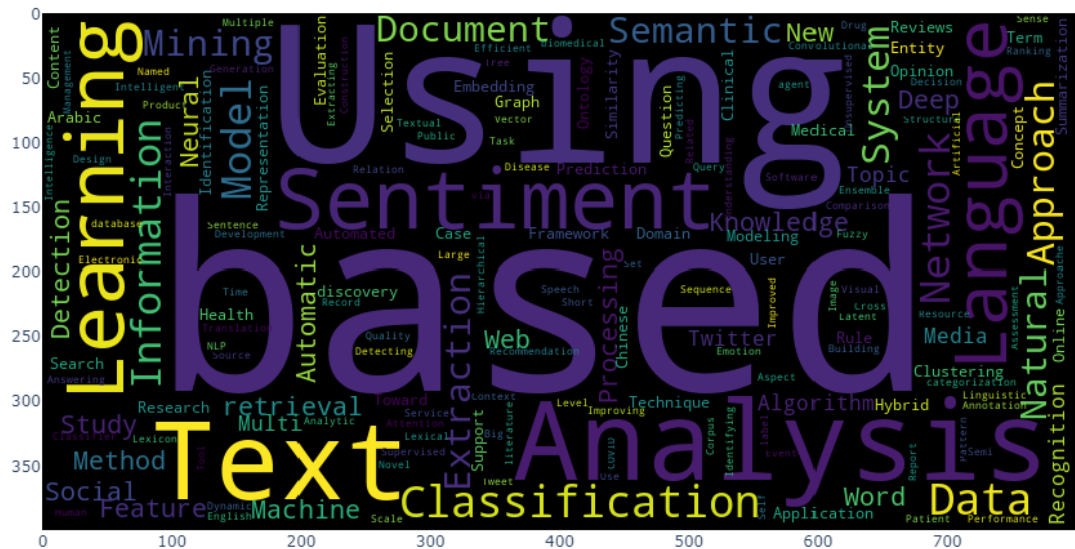


Figure 34. Word Cloud For ‘Article Title’

Word cloud regarding ‘Abstract’ is visualized in Figure 35. Dominant words/phrases are seen as “model”, “data”, “text”, “method”. It seems correlated to “Article Title” in overall. Selected domain “Text Mining” seems evaluated by developed/implemented “model” or “method”. To be able to run that “model” or “method”, “text” and “data” are really required.



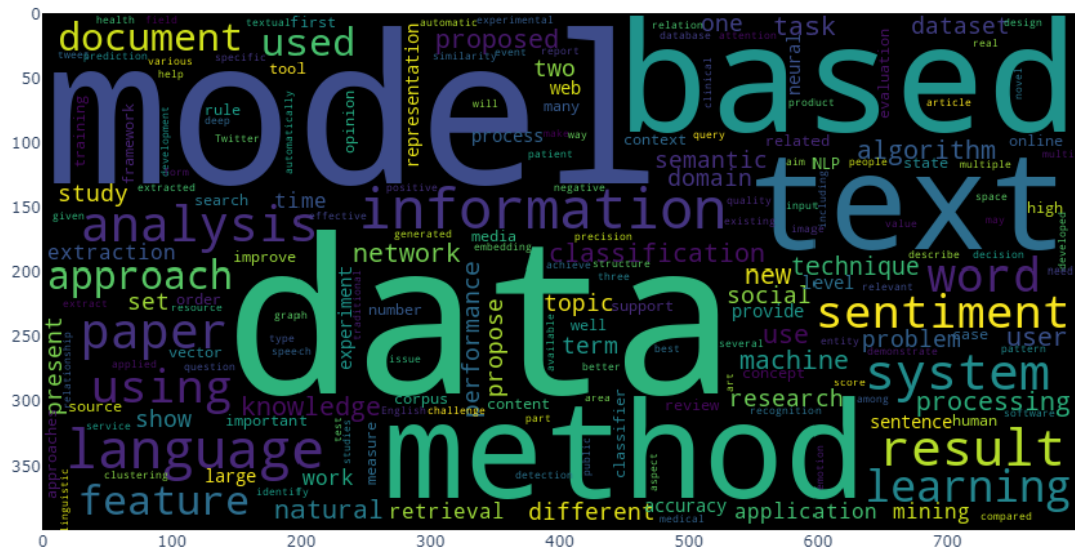


Figure 35. Word Cloud For ‘Abstract’

Word cloud regarding ‘Author Keywords’ is visualized in Figure 36. Dominant words/phrases are seen as “language”, “processing”, “mining”, “learning”, “machine”. They seem reasonable regarding “Natural Language Processing”, “Machine Learning”, “Text Mining” topics. In addition to, there are also other shared words with “Article Title” and “Abstract” such as “Sentiment”, “Semantic”. Furthermore, there are specific keywords such as “network”, “health”.

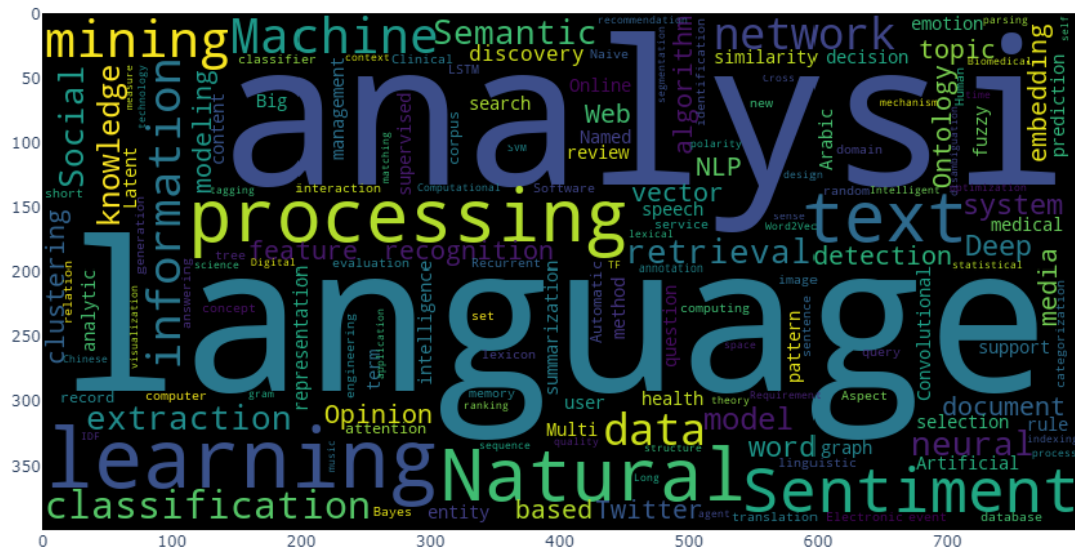


Figure 36. Word Cloud For ‘Author Keywords’

Word cloud regarding ‘Keywords Plus’ is visualized in Figure 37. Dominant words/phrases are seen as “CLASSIFICATION”, “MODEL”, “SYSTEM”, “NETWORK”, “INFORMATION”. The content in ‘Keywords Plus’ are terms or phrases that regularly are mentioned in the references titles but not in the article's title. Regarding this information, having different representation than ‘Author Keywords’ seems reasonable.

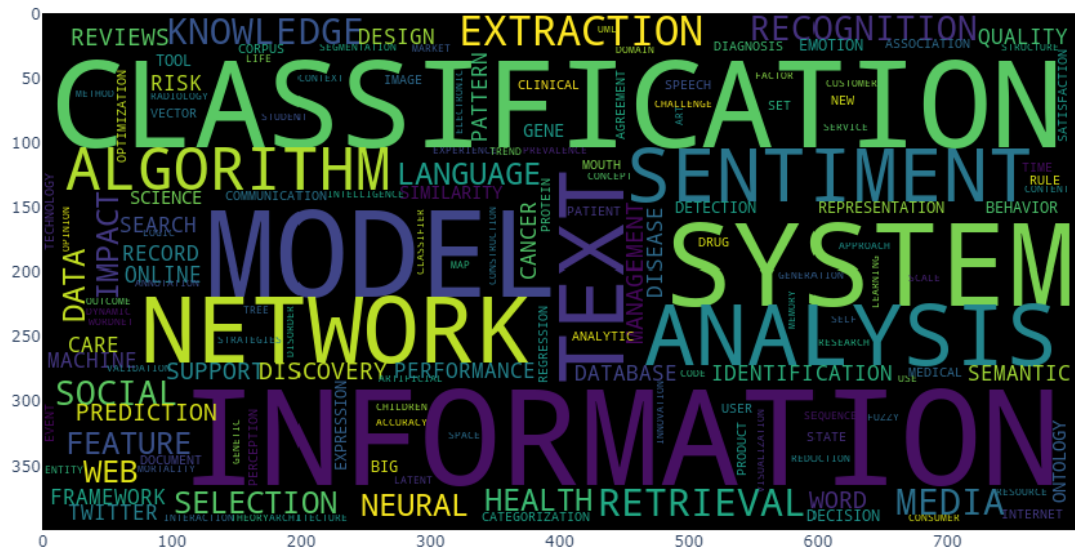


Figure 37. Word Cloud For ‘Keywords Plus’

Word cloud regarding 'AT\_A\_AK\_KP' ( 'Article Title' + 'Abstract' + 'Author Keywords' + 'Keywords Plus') is visualized in Figure 38. For overall combination, dominant words/phrases are seen as “model”, “use”, “data”, “language”, “text”. That is the cumulative representation of overall dataset utilized in this study. It can be seen that main words/terms related to “Text Mining” domain are protected. 'Article Title' and 'Abstract' columns compose first part of the dataset. 'Author Keywords' and 'Keywords Plus' columns compose second part of the dataset.



Generic English language stop words and some additional common words are removed from the content as well. Lemmatization operation is applied to reduce words to their root form. As last step, the words whose length is lower than 3, are excluded.

**3.2.3 Feature extraction.** In the proposed model, majority of the features are in textual form. There are several methods to represent text features to be utilized in Machine Learning, Deep Learning models/algorithms. Rather than traditional BoW techniques, we selected word embedding as an advanced technique.

Word embedding aims to represent the words' semantic and syntactic information. Semantic information here stands for identifying the meaning of words. For referring to structural roles of words, syntactic information is considered. There are different word embedding approaches based on method of matching words to latent spaces (Li & Yang, 2018). In this study, we used one of Neural Network Language Models, Word2Vec to represent the text in our dataset.

Word2Vec proposed by Mikolov et al. (2013) which includes Skip-Gram and CBOW, is very popular in NLP tasks. It can be considered as a two-layer Neural Network Language. In this study, we applied Word2Vec model/technique from scratch. Our methodology is to train Word2Vec model as online learning month by month. Regarding this, we selected 1992 January as it is the time point where there is at least 1 article for every month till 2021 December. To be able to include various N-gram words, word n-gram collocations from a stream of sentences (Bouma, 2009) are applied within Word2Vec. Word2Vec model is trained in online form monthly. For timeframe between 2021 and 1992, totally 360 (30\*12) trained models are generated/saved to be used in next month learning iteration. Dimension of word embedding vector is decided to be 300.

Regarding online Word2Vec training, some statistics about number of vocabulary and N-gram distribution can be seen in the following Table 13.

Table 13  
*Word2Vec Training Statistics*

Period	1-Gram	2-Gram	3-Gram	4-Gram	5-Gram	Total
1995 December	98	129	15	1	1	244
2000 December	387	425	66	4	1	883
2010 December	1518	1124	166	9	2	2819
2015 December	2653	1465	213	16	3	4350
2018 December	4058	1529	221	18	3	5829
2021 December	5995	1550	222	19	3	7789

Sample N-Gram words look like as in the following. 1-Gram (network, knowledge, system, domain, application, information, instruction, brain, signal, rule); 2-Gram (acoustic emission, recall precision, relational database, bilingual memory, signature algorithm, digital system, object-oriented code, intelligent assistant, semantic interpretation, social distance); 3-Gram (logical form generation, dictionary extraction pattern, intelligent data analysis, corporate knowledge repository, feature-based memory association, recurrent neural network, latent semantic analysis, decision support system, multimodal information retrieval, formal concept analysis); 4-Gram (speech natural language processing, concept cluster discover database, genetic algorithm simulated anneal, chinese name entity recognition, cross language information retrieval)

**3.2.4 Model creation.** In this study, forecasting word embedding scenario is approached/evaluated as a time-series problem. Regarding this, generated word embedding vectors are processed as an input through LSTM network model from the point of deep learning view. LSTM is a particular type of RNN, capable of understanding long-term dependencies (Hochreiter & Schmidhuber, 1997).

Details about proposed deep learning architecture is pictured in Figure 39. As a novelty, rather than aggregated document embedding, LSTM network is applied per word individually. We set up LSTM deep learning model to handle the feature set (word embedding matrices) as time-series data. Since word feature is represented with 300 values, we applied multi-variate LSTM methodology. Timestep value is selected as 36 (months) and target feature is processed as the next 36th month.

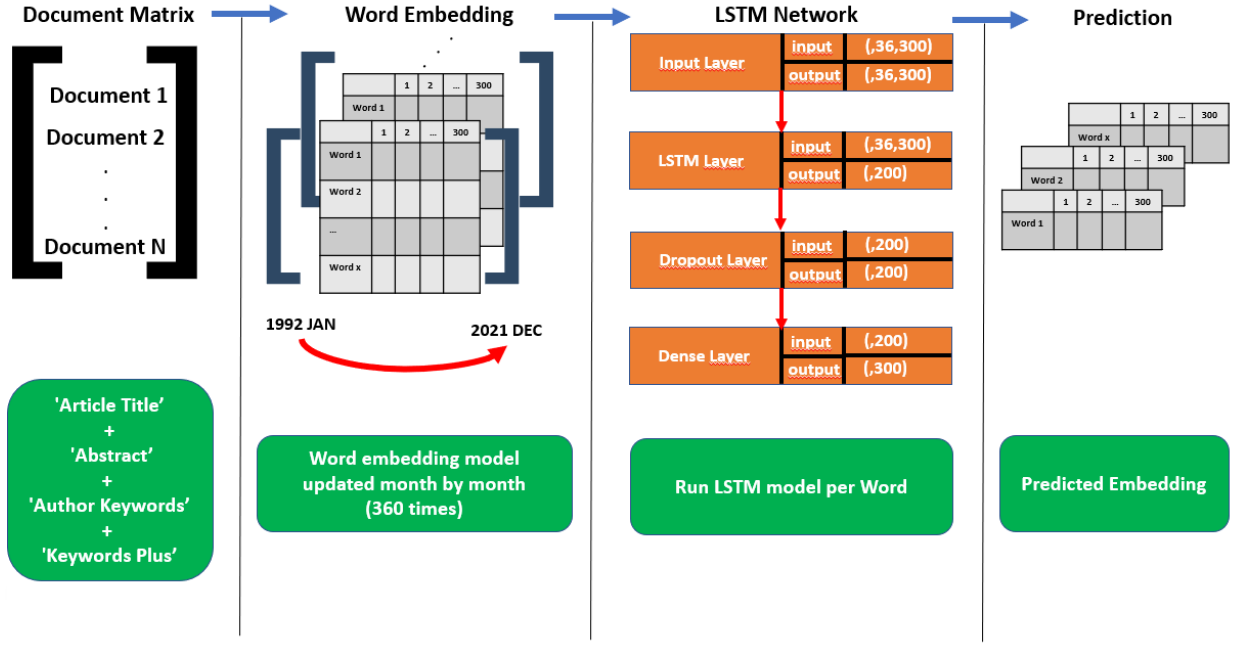


Figure 39. Deep Learning Architecture

For training and validation, we use the articles between 1992 January and 2015 December as input to predict the word embedding matrix of 2018 December. As the timestep value is identified as 36 months, the reference vocabulary is determined as of 2015 December. For testing, the word embeddings from 2016 January to 2018 December are also included in the input and the word matrix of December 2021 is predicted.

There are 5829 distinct words in the vocabulary as of 2018 December before the term selection process. For determining the final vocabulary, first a normalization operation is calculated for each term based on the following formula:

$$\underline{w}_i = w_i / m \quad (4)$$

where  $w_i$  denotes the frequency of  $i^{th}$  term in the vocabulary,  $m$  is the number of months passed since the first occurrence of  $i^{th}$  term, and  $\underline{w}_i$  represents the obtained normalized value of the  $i^{th}$  term for term selection. The normalization operation in Eq. 4 applied before the term selection step is performed in order to put forward the domain-specific terms that have been used in recent years. After calculating the normalized values, median of the normalized values is selected as threshold per N-Gram (1 to 4) separately. As a result, a total of 2944 terms have been selected. These 2944 words are represented with 1914 unique words as of 2015 December. To construct the vocabulary of the test set, the normalization operation given in Eq 4. is applied

for each term obtained as of 2021 December. The distribution according to Word2Vec vocabulary of 2021 December, there are 7789 words. After calculating normalized values, median value is selected as threshold per N-Gram (1 to 4) specific. Regarding this, 3898 words are selected. N-Gram distribution regarding 2018 Test Vocabulary and 2021 Test Vocabulary can be seen in Figure 40.

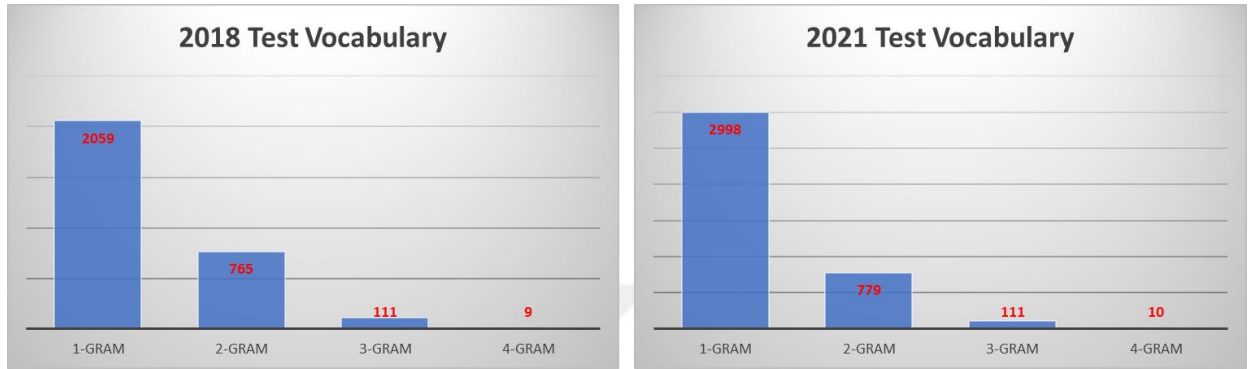


Figure 40. Distribution Of Vocabulary Set Based On N-Gram

**3.2.5 Visualization and quantitative analysis.** For visualization purpose, word embedding values are used to compute cosine similarity between word pairs. Cosine similarity is a metric utilized to compare the documents/words from the point of semantic similarity based on calculation of cosine distance. This similarity metric is more interested in direction than in magnitude. In other words, the similarity between two cosine vectors aligned in the same direction is 1, whereas the similarity between two vectors oriented perpendicularly is 0. When two vectors are diametrically opposed, or pointed in completely opposite directions, the similarity measurement is -1. (i.e., back-to-back). However, Cosine Similarity is frequently used in positive space, between 0 and 1. Cosine Similarity is primarily considering differences in orientation but is only concerned with differences in magnitude/length. Some examples for different vector representation can be seen in Figure 41.



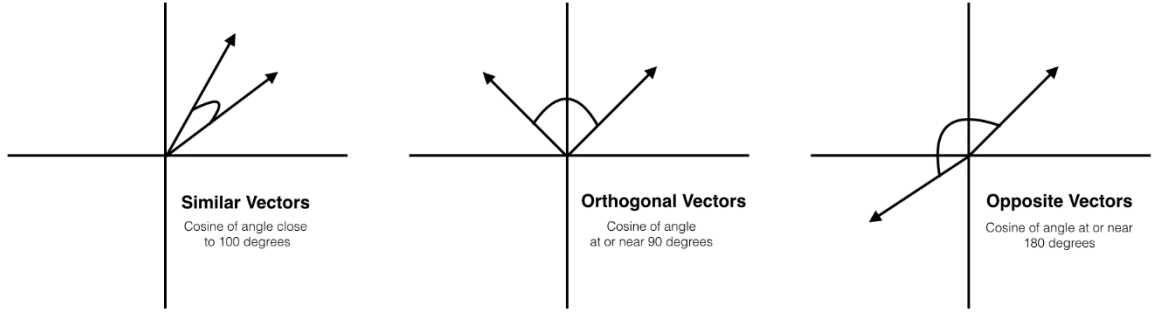


Figure 41. Vector Representation (Tellez Et Al., 2017)

Mathematically, it calculates the cosine of the angle formed by two vectors projected in multiple dimensions. Finding the cosine of the two non-zero vectors is the first step in the Cosine Similarity assessment. The dot product of two Euclidean vectors A and B is defined as in the following formula.

$$A \cdot B = ||A|| ||B|| \cos \theta \quad (5)$$

The cosine similarity is then calculated using the two vectors and the dot product which is shown in following equation. The output will be a number ranging from -1 to 1, with -1 indicating non-similarity, 0 suggesting orthogonality (perpendicularity), and 1 signifying total similarity.

$$similarity = \cos(\theta) = \frac{A \cdot B}{||A|| ||B||} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (6)$$

As Word2Vec model provides word embedding vectors, cosine similarity can be calculated and used for semantic similarity in an efficient way (Jatnika et al., 2019)

For visualization application, we searched alternatives firstly such as Pajek (Batagelj & Mrvar, 2004), VOSviewer (Van Eck & Waltman, 2013). We used Gephi (Bastian et al., 2009) for visualizing purposes regarding cosine similarity matrixes. Gephi is an open-source graph and network analysis program. It uses a 3D render engine to quickly expedite research by displaying large networks in real time. New opportunities to engage with big data volumes and provide useful visual results are made possible by a multi-task and flexible design. There are

major Gephi features in relation to interactive network exploration and comprehension. It offers rapid and easy access to network data and enables spatializing, filtering, traversing, changing, and clustering of network data. There are several literatures works regarding using of Gephi. We used datasets from Dbpedia, Facebook, and Twitter in this paper. Hussain et al. (2018) utilized Gephi and R to investigate the structure of such data and compare various statistics based on the graphs by browsing the graphs. The main focus of the study in (Wajahat et al., 2020) is on the visualization of social networks to better understand community involvement. All the metrics are tailored to certain profiles who engage in media houses and magazine fan sites, as well as likes and favorites. They extracted data from our personal Facebook profile to visualize in the Gephi tool for this investigation. A new framework for effective social network analysis is offered (Thangaraj & Amutha, 2018). The framework is being implemented to track performance. The performance metrics are a significant improvement over the previous methods. A Gephi screenshot can be seen in Figure 42.

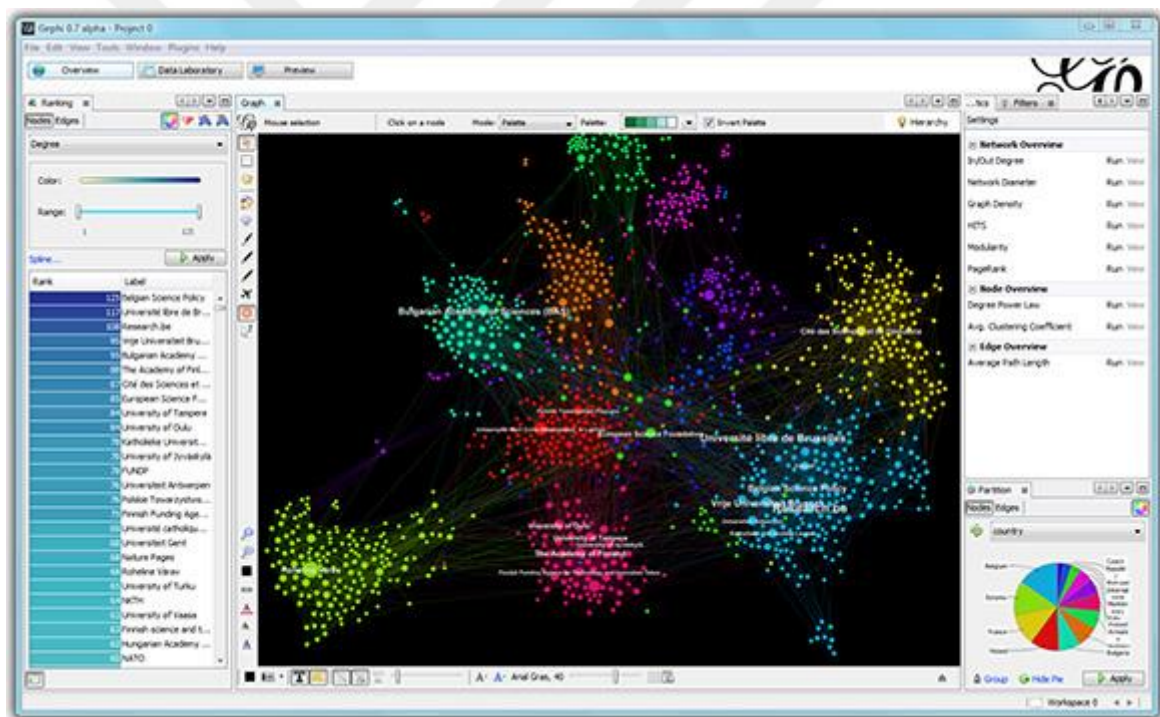


Figure 42. Screenshot Of Gephi

Prior to giving cosine similarity matrix to Gephi, LDA topic modeling method was aimed to be utilized to reduce dimensionality. Following steps are applied for this purpose. However, this method did not help in a good way.

- Calculate cosine Similarity between word pairs for 2017 December Actual, 2020

## December Actual and 2020 December Predict

- Perform Min-Max Scaling
- Reset to zero for diagonals
- Give these inputs to LDA
- Create a new sheet based on Topics and word weights in that Topic
  - For example, we have 5 Topics include 250 words totally (5,250)
  - Apply transpose operation and have (250,5)
  - Dot multiplication (250,5) X (5,250)
  - Output matrix is (250,250)

For the visualisation of the actual and predicted word clusters, the word embedding values are used to calculate the cosine similarity between word pairs. Cosine similarity determines the cosine of the angle formed by two vectors that are projected in three dimensions. As the word2vec model provides word embedding vectors, cosine similarity is a suitable way to represent the semantic similarity between the words of the vocabulary (Jatnika et al., 2019). The matrix including the cosine similarity between the predicted word embedding vectors for the test set representing 2021 December is given as input to force-atlas2 algorithm (Jacomy et al., 2014). The same procedure is also applied to the actual word embedding vectors obtained using the published documents as of 2021 December. The obtained predicted and actual word clusters are displayed and analysed in comparison with each other. Regarding measuring the performance of clustering evaluation, several metrics are used. The Rand Index (RI) is a measure of the similarity between two clusters in term of statistics. The Adjusted Rand Index (ARI) is the updated version with chance corrections of the Rand Index (Hubert & Arabie, 1985). It is expressed in the following formula:

$$ARI = (RI - Expected\_RI) / (Maximum\_RI - Expected\_RI) \quad (7)$$

Adjusted Mutual Information (AMI) score is an adjustment of the Mutual Information (MI) score to account for chance (Vinh et al., 2010). It explains why the MI for two clustering results with a larger number of clusters is generally higher, regardless of whether more information is really shared. For two clustering evaluation U and V, the AMI is expressed in

the following formula:

$$AMI(U, V) = [MI(U, V) - E(MI(U, V))] / [avg(H(U), H(V)) - E(MI(U, V))] \quad (8)$$

Homogeneity score is a metric of a cluster labelling given a ground truth. If all of a clustering result's data points are members of one class, then that clustering result is homogeneous (Rosenberg & Hirschberg, 2007). This score is expressed in the following formula:

$$H = 1 - \frac{H(Y_{true}|Y_{pred})}{H(Y_{true})} \quad (9)$$

Completeness score is a metric of a cluster labelling given a ground truth. If every data point that belongs to a certain class is a part of the same cluster, the clustering result is complete (Rosenberg & Hirschberg, 2007). This score is expressed in the following formula:

$$C = 1 - \frac{H(Y_{pred}|Y_{true})}{H(Y_{pred})} \quad (10)$$

V-measure score is identical to normalized mutual information score using the arithmetic averaging option (Rosenberg & Hirschberg, 2007). This score is expressed in the following formula:

$$V = (1 + beta) * H * C / (beta * H + C) \quad (11)$$

Cosine similarity matrix for 2018 December Actual consists of 2944X2944 cells. In Figure 43, we aimed to display a subset. In this representation, it can be evaluated that for example “model” and “method”; “classification and result” have a strong relationship.

	analysis	model	system	method	base	information	approach	classification	result	paper
analysis	0	2.974176501	3.377169248	3.634421167	3.523608784	3.173696041	3.47148786	3.53724207	3.423826778	4.499163719
model	3.305186264	0	4.076143676	5	3.900568396	3.007452475	5	4.142807361	4.237743519	2.303826304
system	3.342313824	3.590649605	0	3.427119147	2.786454017	4.475676414	3.045078024	2.517734174	2.005779085	2.546721764
method	4.172313186	5	3.929680155	0	4.218251199	2.857275161	4.881316106	4.041440432	4.427528361	3.223441734
base	3.796716608	3.683076524	3.078285047	3.920395364	0	3.570635381	3.425472221	3.22631246	2.814931286	3.729566647
information	2.929358114	2.54148491	4.207247996	2.354641727	3.065647234	0	3.287791417	2.570664112	2.248078074	3.118547776
approach	3.958828521	4.937824711	3.487890692	4.821452982	3.627462937	4.000404685	0	2.528142653	3.952828633	3.658040065
classification	4.223494661	4.335179375	2.966663864	4.210266966	3.549567754	3.225921457	2.591887115	0	4.446206656	2.926213806
result	4.031902344	4.377998158	2.332079736	4.558016176	3.013838857	2.791693989	4.100338491	4.388231801	0	2.877282102
paper	4.89119038	2.373897276	2.873338234	3.109421904	3.788917457	3.67527756	3.503809061	2.755635302	2.735710952	0

Figure 43. Scaled Cosine Similarity Matrix For 2018 December Actual

Cosine similarity matrix for 2021 December Actual consists of 2944X2944 cells. In Figure 44, we aimed to display a subset. Again, “model” and “method”; “classification and result” have a strong relationship. Let us focus on cosine similarity values for “analysis – model” and “analysis – system”. Cosine similarity of “analysis – model” increased in 2021

December Actual while “analysis – system” decreased. Finally, cosine similarity of “analysis – model” is greater than “analysis – system”

	analysis	model	system	method	base	information	approach	classification	result	paper
analysis	0	3.26815612	2.334726036	3.75586358	4.381278101	3.639195455	4.294772345	3.714373911	3.333856809	3.651369465
model	3.936571729	0	4.148236517	5	3.600787997	2.848201756	4.860578816	4.180138466	4.171692901	2.874856004
system	2.436567125	3.487121991	0	3.819235987	2.523767429	3.500746413	4.598284425	2.817245927	3.088615585	2.964268645
method	4.077484168	4.491011651	4.090056612	0	2.727528405	4.188773156	5	3.954144292	4.176636329	2.994080985
base	4.728588948	3.266772003	2.748623161	2.802585534	0	3.373473437	3.118687437	3.568897528	3.078064305	3.535229529
information	3.890270513	2.595685383	3.688910213	4.078104623	3.302255727	0	3.243929615	2.676563714	1.555333844	3.522379608
approach	4.657592546	4.332963792	4.930969922	4.95293599	3.140397864	3.334708805	0	3.566437022	3.152634943	3.222830286
classification	4.403240186	4.11214095	3.248708019	4.323307747	3.889671757	2.887640677	3.868256919	0	4.066108664	3.623074428
result	4.036984914	4.217934515	3.661392414	4.698815838	3.384826391	1.494517744	3.456786029	4.176610037	0	3.149709841
paper	4.128759231	2.736378619	3.31057337	3.146882218	3.668121846	3.741002769	3.315568302	3.460189977	2.97616856	0

Figure 44. Scaled Cosine Similarity Matrix For 2021 December Actual

Cosine similarity matrix for 2021 December Predict consists of 2944X2944 cells. In Figure 45, we aimed to display a subset. If we look at cosine similarity values for “analysis – model” and “analysis – system”, we can say that cosine similarity of “analysis – model” is greater than “analysis – system” as in 2021 December Actual.

	analysis	model	system	method	base	information	approach	classification	result	paper
analysis	0	2.722804124	1.920335872	1.780505166	4.184203091	1.347747813	1.936524415	1.952759339	2.415088154	1.817102938
model	2.085656551	0	1.747672814	1.760430689	4.671984222	1.60818339	1.372525164	2.464506808	2.601577811	2.174480411
system	2.453473122	2.956941011	0	1.537603674	3.383169119	1.720832463	2.117905816	1.936375829	1.998445299	1.901708305
method	2.254015508	2.963661732	1.538064391	0	2.516109987	1.349475439	2.185981446	2.236006498	3.760472965	1.874926194
base	3.143027365	5	2.075876143	1.687462397	0	2.047414141	1.574455543	2.169897254	2.722822004	1.974067485
information	1.569946338	2.351384044	1.572804144	1.241116414	2.891843142	0	1.583158703	2.102720814	2.220868249	1.422013572
approach	2.203551113	2.059521569	1.855908804	1.919540911	2.083218667	1.623904589	0	2.427451378	2.54309234	1.786771906
classification	1.875764752	3.545732433	1.467563325	1.787065539	2.575330551	1.798048073	2.289314721	0	3.549420886	2.32916802
result	2.009335062	2.85627591	1.323000826	2.474399964	2.730962412	1.59244189	1.840122863	2.743931927	0	1.848929076
paper	2.087934485	3.417111284	1.717736628	1.701613907	2.734171065	1.404976657	1.770723736	2.538556628	2.553081409	0

Figure 45. Scaled Cosine Similarity Matrix For 2021 December Predict

Cosine similarity matrix for 2024 December Predict consists of 3898X3898s cells. In Figure 46, we aimed to display a subset.

	model	analysis	method	system	base	information	approach	classification	result	covid-
model	0	1.71537665	1.754496744	1.900967771	2.598956152	1.687842188	1.926213398	2.93921616	2.175268876	1.590338648
analysis	1.773738549	0	1.904101627	1.647705946	2.284978822	1.37099041	1.683478819	3.708131141	2.115427618	2.341676173
method	1.783407613	1.874445	0	2.077704395	2.847995951	1.629818572	2.010933059	3.299659466	1.797747308	1.68546043
system	1.922885853	1.608062821	2.081090751	0	1.73269521	2.10755129	1.815135916	2.500091628	1.762179777	1.888101528
base	1.890355399	1.63660745	2.023095167	1.335882045	0	1.380674805	1.642072221	3.890236854	2.08899883	1.864046048
information	1.738627368	1.366172824	1.650745137	2.119546829	1.872931073	0	1.774585811	3.467882667	1.731957582	1.657257579
approach	1.98330093	1.696751681	2.042912006	1.857231914	2.31685305	1.794557365	0	3.693831843	1.543836189	1.441288933
classification	1.833446445	2.067981797	1.943091643	1.650111842	3.075271761	1.989201661	2.062491854	0	1.774763956	1.96197073
result	1.996898476	1.888037781	1.614211134	1.592899585	2.642367148	1.529552944	1.317677488	2.376154714	0	2.06471184
covid-	1.550943974	2.226210887	1.608497114	1.814105977	2.473120226	1.559082728	1.31311068	3.138229049	2.184045951	0

Figure 46. Scaled Cosine Similarity Matrix For 2024 December Predict

### 3.2.6 Results. To perform the experiments, we used the supercomputer environment.

The computing tools utilized in this work were made available by the National Center for High Performance Computing of Turkey (UHeM). A screenshot from server hardware regarding UHeM can be seen in Figure 47.



*Figure 47. Hardware Of Supercomputer Server (UHeM, 2022)*

List of major Python packages/libraries and their version used in this study are listed in Table 14 as reference.

Table 14  
*Python Package List*

<b>Package</b>	<b>Version</b>
beautifulsoup4	4.10.0
bs4	0.0.1
gensim	3.8.3
Keras	2.4.3
matplotlib	3.4.1
networkx	2.5.1
nltk	3.5
numpy	1.19.2
pandas	1.2.3
pip	19.2.3
pyLDAvis	3.2.2
regex	2021.3.17
scikit-learn	0.24.1

scipy	1.6.2
seaborn	0.11.2
tensorflow	2.4.1
XlsxWriter	1.4.3

Pandas library is used from the point of data analysis and preparation (Team, 2020; McKinney, 2010). NLTK is used for text processing regarding tokenization, lemmatizing and similar operations (Bird et al., 2009). Gensim library (Rehurek & Sojka, 2010) is used to obtain Word2Vec model. We used Keras library with Tensorflow backend to build LSTM models (Chollet, 2021).

For hyper-parameter optimization and fine tuning regarding Word2Vec models, dimensionality of the word vectors, minimum word frequency and vocabulary trimming rule are applied. Regarding LSTM models, model checkpoint and early stopping mechanisms are implemented with cross validation. Validation loss score is selected as early stopping criteria to save best model during training phase. To evaluate performance of LSTM model, Mean Squared Error metric is calculated.

**3.2.6.1 Experimental setup.** To perform the experiments, we used a super-computing environment. The size of the training set consisting of 45406 documents is 150 MB. The trials were performed with a hardware setup consisting of an Intel® Xeon® Gold 6148 CPU (28 Cores, 126 GB Disk) and a NVIDIA Tesla V100 SXM2 GPU. Gensim library (Rehurek & Sojka, 2010) is used to train the Word2vec model on the dataset consisting of the documents from the selected domain. For hyper-parameter optimization and fine tuning of the Word2vec models, dimensionality of the word vectors, minimum word frequency and vocabulary trimming rule are applied. For dimensionality of word vectors, following values are tested: 100, 300. For minimum word frequency, frequency of Unigram words, Bigram words and Trigram words are utilized. A trimming rule feature of Word2vec model is also utilized. This trimming rule is used to decide whether certain words should remain in the vocabulary or should be discarded regarding frequency. Regarding the LSTM model built to predict the future word embedding matrix, we use Keras library with Tensorflow backend (Chollet, 2021). The model checkpoint and early stopping mechanisms are implemented with cross validation. For cross validation, train set is divided as 80% and 20% for training and validation. In addition to, LSTM layers, number of LSTM layers

per unit, epochs, optimizer and loss score parameters are validated, the validation loss score is selected as the early stopping criteria to determine the best model during the training phase. To evaluate performance of LSTM model, mean squared error metric is used. Dimension of the word embedding vector is decided to be 300. Optimal values for overall hyperparameters for the Word2Vec and LSTM model can be seen in Table 15.

Table 15  
*Optimal values of the hyperparameters for the Word2Vec and LSTM Model*

Hyperparameter	Value
Dimension of Word2Vec Vector	300
LSTM Layers	1
Number of LSTM Units Per	200
Epochs	100
Optimizer	Adam
Loss	Mean Squared Error

**3.2.6.2 Quantitative results.** Table 16 shows the quantitative results in terms of the metrics detailed in Section 3.4. We focused on metrics regarding performance of LSTM model and cluster creation. It also illustrates the prediction results for 2021 and performance metrics based on different prediction horizons. For example, the first column “2018” shows the prediction that is performed using the data until 2018 for 2021.

For 2021 December Actual, 37 clusters are identified based on cosine similarity matrices of word pairs. We are able to predict 35 of 37 clusters (95%). This is very promising result. However, ratio of word similarity (62%) is lower than cluster similarity. This points out that there exist different words in the matched clusters. During constructing Word2Vec vocabulary and identifying training vocabulary set, it is ensured that volume of Unigram words does not suppress multi N-Gram words. Regarding this, word similarity for Fourgram, Bigram and Trigram is higher than Unigram.

Results for predicting 2021 December from 2015 and 2012 horizons are worse than 2018 as expected. It reveals that current proposed model outperforms better with short-term prediction (+3 years). However, Overall Matching Clusters statistic still gives promising result for long-term prediction (+6 years and +9 years).



Table 16  
Performance Metric Comparison Using Different Prediction Horizons for 2021

Predicted From Metric	2018	2015	2012	2018 Baseline	2015 Baseline	2012 Baseline
Average Train Mean Squared Error	0.02	0.05	0.09	N/A	N/A	N/A
Average Test Mean Squared Error	0.38	0.86	0.99	N/A	N/A	N/A
Overall Matching Clusters	92%	81% (31/38)	65% (25/38)	89% (34/38)	84% (32/38)	58% (22/38)
Directly Matching Clusters	68%	55% (21/38)	52% (20/38)	68% (26/38)	63% (24/38)	50% (19/38)
Overall Word Similarity	62%	56% (713/1253)	52% (659/1253)	71% (893/1253)	41% (508/1253)	27% (341/1253)
Unigram Word Similarity	60%	57% (553/960)	53% (515/960)	69% (668/960)	35% (338/960)	24% (227/960)
Bigram Word Similarity	71%	53% (132/249)	48% (120/249)	77% (193/249)	58% (145/249)	38% (94/249)
Trigram Word Similarity	68%	63% (26/41)	53% (22/41)	71% (29/41)	59% (24/41)	49% (20/41)
Fourgram Word Similarity	100% (3/3)	66% (2/3)	66% (2/3)	100% (3/3)	33% (1/3)	0% (0/3)
Adjusted Rand Index Score	0.55	0.21	0.14	0.74	0.58	0.43
Adjusted Mutual Information Score	0.65	0.45	0.38	0.78	0.68	0.62
Homogeneity Score	0.69	0.5	0.42	0.77	0.72	0.66
Completeness Score	0.75	0.61	0.58	0.86	0.81	0.79
V-Measure Score	0.71	0.55	0.48	0.81	0.76	0.72

For word “health information”, training and validation loss graphic can be seen in Figure 48 as a sample result.

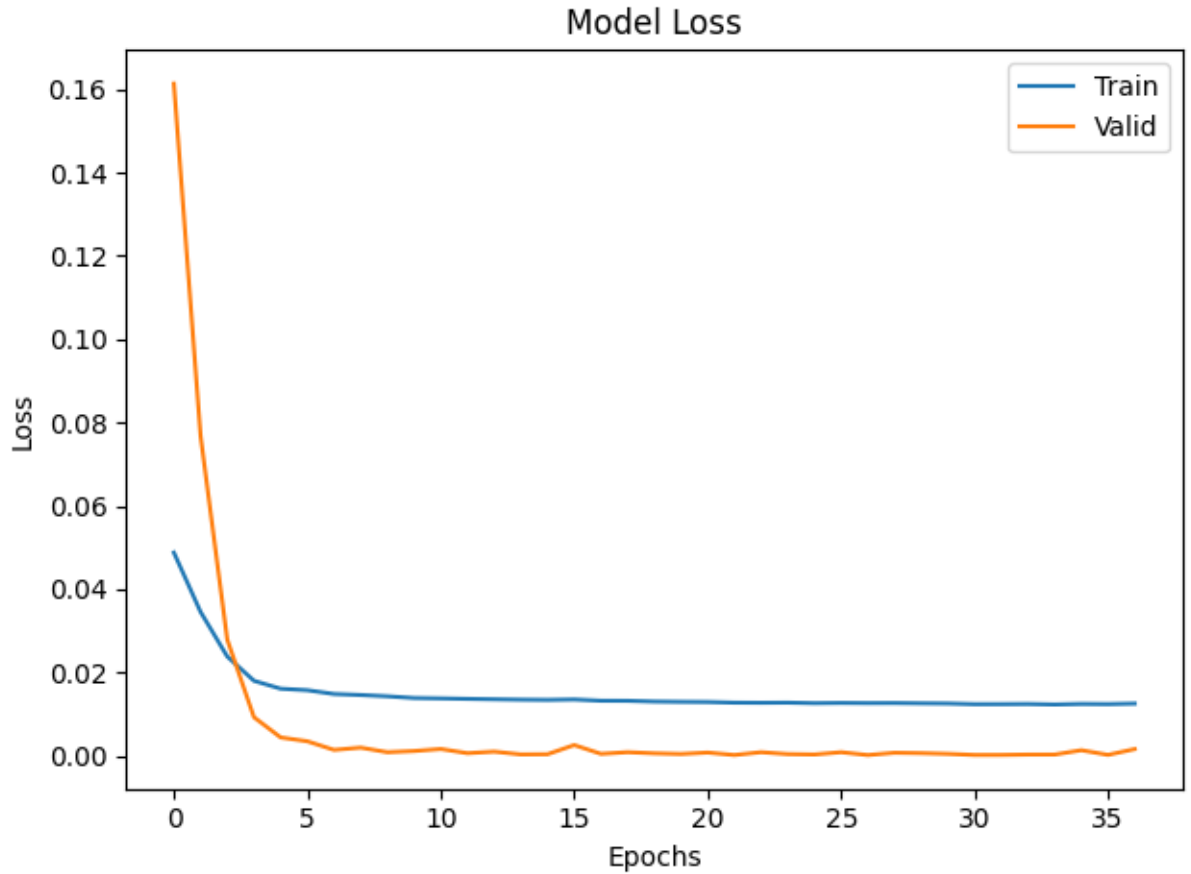


Figure 48. Loss Versus Epoch Performance

**3.2.6.3 Practical results.** Table 17 displays the comparison of clusters for 2018 December Actual, 2021 December Actual, 2021 December Prediction, 2024 December Prediction, 2027 December Prediction and 2030 December Prediction. It can be seen that our proposed TF model successfully works in practice when its performance is examined by matching the actual and predicted clusters. Looking at the predictions for 2024, 2027 and 2030, our model can able to identify new clusters that were not apparent in the past and the model also forecasts that some areas would have a lower popularity in the future as they disappeared in the table. These new clusters represent emerging areas/topics regarding text mining use-case domain.

Table 17  
Comparison of Clusters

Cluster	2018 Dec Act.	2021 Dec Act.	2021 Dec Pred.	2024 Dec Pred.	2027 Dec Pred.	2030 Dec Pred.
C1 – Recommender Systems	Yes	Yes	No	No	No	No
C2 – Lexicon Analysis	Yes	Yes	Yes	Yes	No	No
C3 – Multi-Domain Learning	Yes	Yes	No	No	No	No
C4 – Machine Learning Models/Types	Yes	Yes	Yes	Yes	*Yes	*Yes
C5 – Speech Recognition	Yes	Yes	*Yes	*Yes	Yes	*Yes
C6 – Ontology Learning	Yes	Yes	Yes	*Yes	Yes	*Yes
C7 – Deep Learning	Yes	Yes	Yes	Yes	*Yes	*Yes
C8 – Machine Translation	No	Yes	Yes	Yes	*Yes	*Yes
C9 – Communication	*Yes	Yes	*Yes	*Yes	*Yes	*Yes
C10 – Opinion Mining	*Yes	Yes	Yes	*Yes	*Yes	*Yes
C11 – Concept Analysis	*Yes	Yes	*Yes	*Yes	No	No
C12 – Machine Learning Performance Analysis	Yes	Yes	Yes	Yes	Yes	Yes
C13 – Entity Recognition	*Yes	Yes	*Yes	Yes	Yes	Yes
C14 – Information Extraction	*Yes	Yes	*Yes	No	No	No
C15 – Mental Health	No	Yes	Yes	Yes	Yes	*Yes
C16 – Cyber Security	*Yes	Yes	*Yes	*Yes	*Yes	*Yes
C17 – Topic Modelling and Clustering	Yes	Yes	Yes	Yes	Yes	Yes
C18 – Adverse Drug Events	Yes	Yes	Yes	Yes	Yes	*Yes
C19 – Conversational Agents	Yes	Yes	*Yes	*Yes	Yes	*Yes
C20 – Software Engineering	*Yes	Yes	*Yes	*Yes	*Yes	*Yes
C21 – Image Mining	Yes	Yes	*Yes	*Yes	*Yes	Yes
C22 – Education	Yes	Yes	Yes	Yes	Yes	Yes
C23 – Radiology	*Yes	Yes	*Yes	Yes	Yes	No
C24 – Customer Experience	Yes	Yes	Yes	Yes	Yes	Yes
C25 – Performance Computing	Yes	Yes	Yes	No	No	No
C26 – Music	No	Yes	*Yes	*Yes	*Yes	Yes
C27 – Decision Support Systems	Yes	Yes	Yes	No	No	No
C28 – Word Embedding	No	Yes	Yes	Yes	No	No
C29 – Health	*Yes	Yes	*Yes	Yes	Yes	*Yes
C30 – Biochemistry	Yes	Yes	Yes	Yes	Yes	Yes
C31 – Source Code Analysis	*Yes	Yes	*Yes	*Yes	*Yes	*Yes
C32 – Web Security	*Yes	Yes	*Yes	*Yes	*Yes	*Yes
C33 – Knowledge Management	*Yes	Yes	*Yes	No	No	No
C34 – Language Translation	Yes	Yes	Yes	*Yes	*Yes	*Yes
C35 – Social Network Platforms	*Yes	Yes	*Yes	*Yes	*Yes	*Yes
C36 – Finance	Yes	Yes	Yes	Yes	Yes	Yes
C37 – Requirement Engineering	*Yes	Yes	*Yes	Yes	*Yes	*Yes
**Pandemic	No	No	No	Yes	*Yes	Yes
**Disaster Management	No	No	No	Yes	*Yes	*Yes
**Sustainable Energy	No	No	No	Yes	*Yes	*Yes
**Digital Entrepreneurship	No	No	No	No	Yes	*Yes
**Smart Transportation	No	No	No	No	Yes	*Yes
**Smart Devices/IoT	No	No	No	No	No	Yes
**Business Management	No	No	No	No	No	Yes

The cells include (\*) means that these clusters do not exist directly in the regarding case however they are included/grouped under a similar cluster. The cells include (\*\*) means that they only exist for 2024 Dec Predict, 2027 Dec Predict and 2030 Dec Predict.

In the following Figure 49, 50, 51, 52, 53 and 54, the calculated cosine similarity of word pairs is visualized with the help of Gephi (Bastian et al., 2009) regarding the identified cluster names as shown in Table 17. Existing clusters are marked with circle form while newly identified clusters are marked with rectangle form.

In Figure 49, all of the identified 37 clusters are visualized. Name of the clusters can be seen in Table 17. There are several clusters regarding using chosen text mining area. The

clusters consist of both AI, machine learning areas (C4, C7, C12, C17 etc.) and application areas (C5, C16, C29, C36 etc.). For predictions of 2024, 2027 and 2030, the model identified new clusters as emerging areas/topics which are very promising. In addition to discovering new emerging areas/topics, we also observe appearance of new words inside existing clusters. This is also very promising to realize the change for an existing cluster come from past.

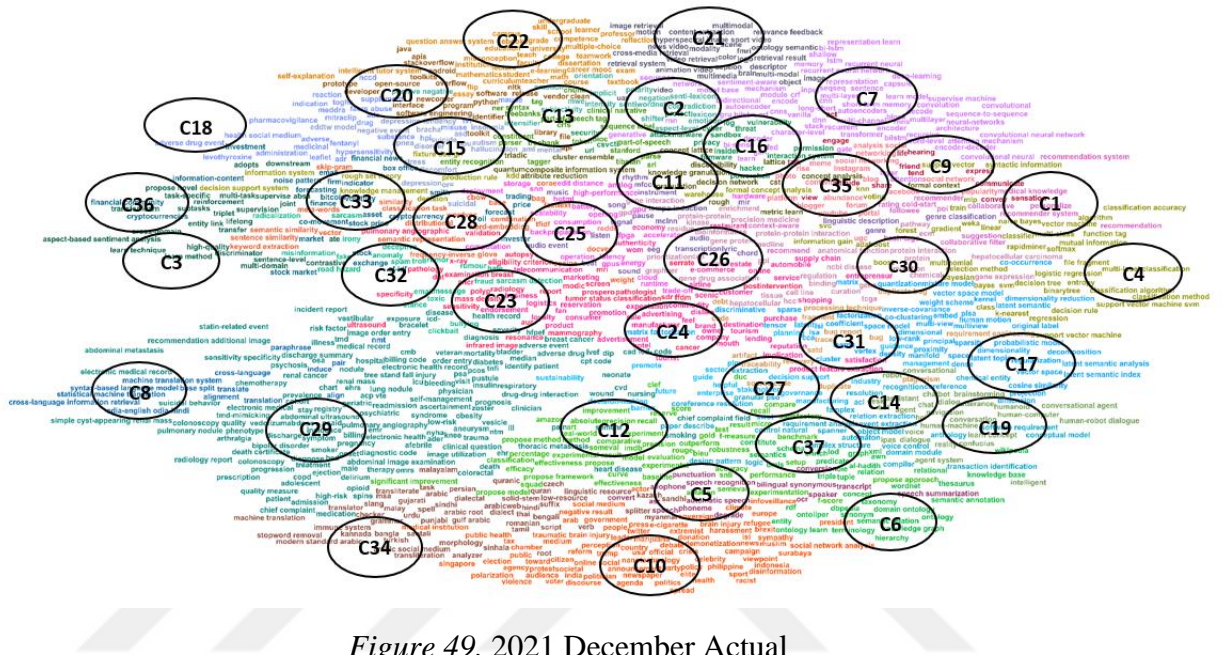


Figure 49. 2021 December Actual

In Figure 50, actual Word2Vec embedding vectors for 18 December month is displayed. According to a comparison between 2018 December Actual and 2021 December Actual, there seem to be 4 new clusters as emerging and these are: “Mental Health”, “Machine Translation”, “Music” and “Word Embedding”. Our prediction model managed to identify all

of these 4 clusters in the results of 2021 December Predict as shown in Table 17.

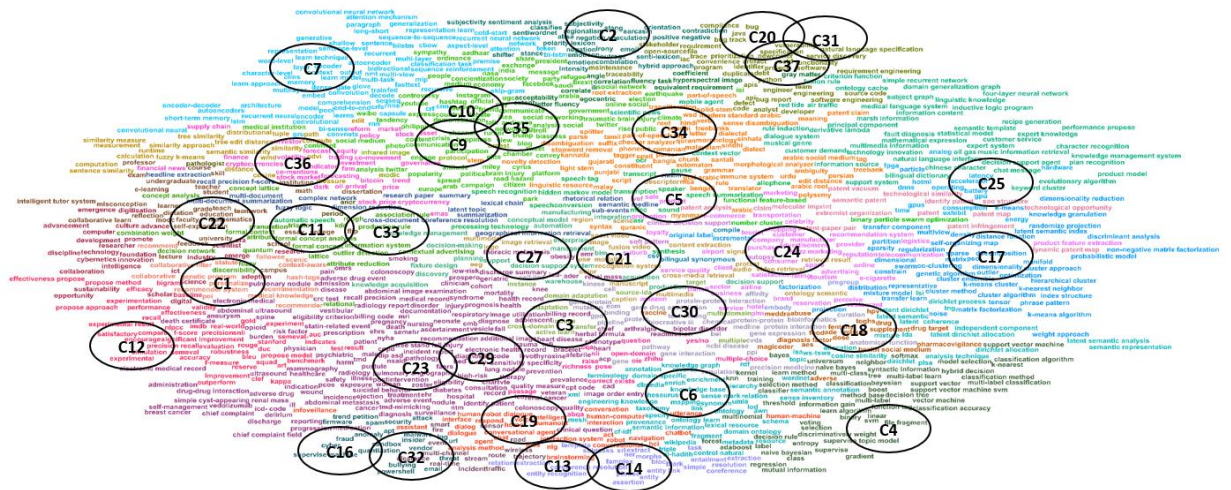


Figure 50. 2018 December Actual

In Figure 51, predicted Word2Vec embedding vectors for 2021 December month is displayed. Regarding comparing 2021 December Predict and 2021 December Actual, 35 clusters are matched and 2 clusters are different only. Looking at exact number of directly matched clusters, there are 26 clusters and 8 clusters are overlapping in the 2021 December Predict results where there are areas belong to multiple clusters and merged into single clusters. These overlapping clusters are: Entity Recognition and Information Extraction; Concept Analysis and Knowledge Management; Social Network Platforms and Communication; Cyber Security and Web Security; Health and Radiology; Image Mining and Music; Requirement Engineering, Software Engineering and Source Code Analysis; Conversational Agents and



## Speech Recognition.

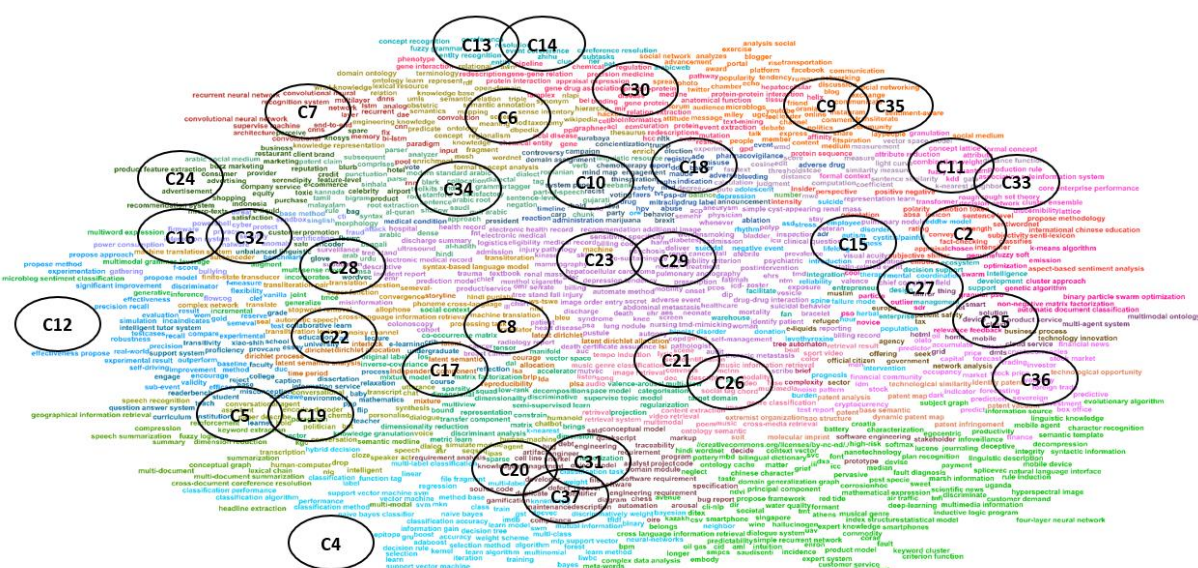


Figure 51. 2021 December Predict

In Figure 52, predicted Word2Vec embedding vectors for 2024 December month is displayed. Based on comparison with 2021 December Actual, it can be seen that 31 clusters still exist in the future. There are identified 3 new clusters as emerging areas/topics. These clusters are “Pandemic”, “Disaster Management” and “Sustainable Energy”. Let’s talk and discuss about the rationale for these new emerging areas/topics.

- “Pandemic” (Subset of Vocabulary: pandemic, covid, influenza, vaccine, coronavirus)** - An unforeseen and dangerous situation (COVID-19 disease) that threatened all of humanity occurred at the beginning of 2020. During this period, several studies are submitted into literature to apply especially machine learning and deep learning techniques for different aspects such as analyzing people emotions and opinions from social network, predicting virus mutation etc. Between June 15 and November 15, 2020, about 300 million online data sources were analyzed using NLP algorithms to find COVID-19-related and irrelevant subjects (Massey et al., 2021). In a similar social media analysis, Cabezas et al. (2021) used a total of 3 million tweets to detect emotional change during the COVID-19 pandemic. To better comprehend public attitudes, worries, and feelings that could influence the achievement of herd immunity objectives, Lyu et al. (2021) aimed to identify the topics and sentiments in the public conversation on social media about the COVID-19 vaccination as well as to identify significant shifts in themes and opinions over time. Mohamed et al. (2021)

used the seq2seq LSTM neural network to predict next-generation sequences of RNA virus while treating these sequences as text data. Du et al. (2021) aimed to create and assess an intelligent automated approach for spotting and categorizing false material on the human papillomavirus (HPV) vaccine on social media using machine learning-based techniques.

- **“Disaster Management” (Subset of Vocabulary: disaster, hurricane, earthquake)**
  - Regarding global climate challenges, the disaster scenarios become more visible. Social media's enormous data output offers an important opportunity for disaster research. Due to the limitations of employing physical sensing technologies, such as the demand to cover a broad region in a short period of time, utilizing of social sensing gains importance. Because of word ambiguity, it might be challenging to differentiate between disaster tweets from typical ones. Song and Huang (2021) offer a sentiment-aware contextual model for Tweet-based catastrophe detection. Taking into account how a post makes you feel about a disruption—whether it's reporting a real disturbance (negative), a disruption in general, or not being influenced by a disruption at all, Roy et al. (2020) offer a multilabel classification strategy to identify the co-occurrence of several types of infrastructure disruptions (positive). Additionally, they provide a dynamic mapping system for displaying interruptions to infrastructure. Sakahira and Hiroi (2021) suggest a technique for assessing cascading disasters that uses NLP to completely and impartially uncover causal relationships between disaster events based on news stories and create networks of cascading disasters.
- **“Sustainable Energy” (Subset of Vocabulary: energy, carbon, grid)** - From the point of sustainability, using energy resources efficiently and investing renewable energy resources are very important. The current global energy crisis has made it even more urgent to move toward clean energy and has once again brought attention to the crucial role that renewable energy plays. Building a charging infrastructure is a good way to encourage the adoption of new energy cars. Wang et al. (2021) examines consumer preferences for charging infrastructure based on comments made by consumers on open social media platforms using NLP technologies. Lannelongue et al. (2021) measures to contextualize greenhouse gas emissions are developed and a methodological framework for standardizing and reliably estimating the carbon footprint of every computational operation is offered. The short-text descriptions of the secondary electrical equipment breakdowns have received a lot of attention. These

failures have happened as a result of the complete development of the smart grid and the substantial operational data generated by the functioning of the power grid. Wei et al. (2020) examines the fault data that arises from the secondary equipment's operation.

For C29 – Health cluster, **“infodemiology”** word appears in the cluster. Pollack et al. (2021) study on a longitudinal infodemiology work to analyse the changes in telemedicine's terminology and attitude during the COVID-19 epidemic. For C16 – Cyber Security, **“cybercrime”** and **“cyber-attack”** appear in the cluster. Due to the integration of heterogeneous devices with varying degrees of processing, communication, and power capabilities, cyber-physical systems provide a complicated array of security concerns (Datta et al. 2020). Users of the public cloud, whether they are potential cyberattack victims, cybercriminals, or digital forensic investigators, naturally converse using words and semantics in document messaging like texts, emails, and instant conversations. As a result, communication using genuine human language gives cloud users a unique identify (Baror et al., 2021). For C28 – Word Embedding, **“mobilebert”** word appears in the cluster. Recent advances in NLP have been made thanks to the use of massive pre-trained models with hundreds of millions of parameters. These models can't be applied to mobile devices with limited resources because of their large model sizes and excessive latency (Sun et al., 2020). For overlapping of C9 – Communication, C10 – Opinion Mining and C35 – Social Network Platforms, **“influencer”** word appears in the cluster. The term "virtual influencer" has evolved and is gaining popularity in social media and network services due to the quick development of social-commerce (Park et al., 2021).



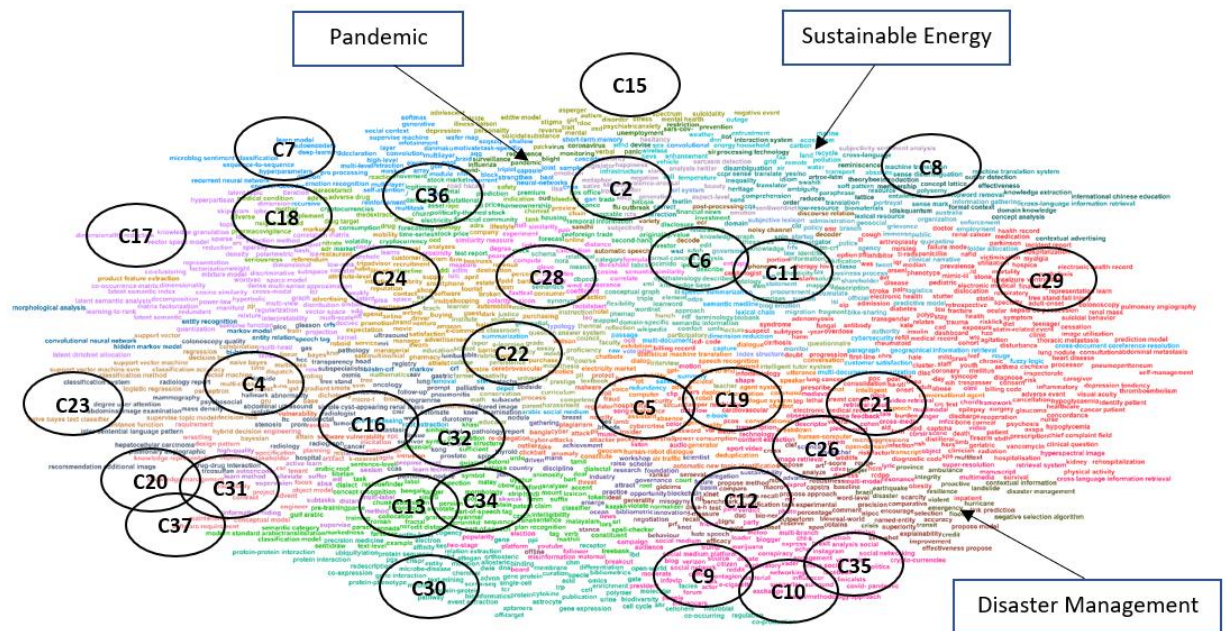


Figure 52. 2024 December Predict

In Figure 53, predicted Word2Vec embedding vectors for 2027 December month is displayed. Based on comparison with 2021 December Actual, it can be seen that 28 clusters still exist in the future. There are identified 2 new clusters as emerging areas/topics. These clusters are “Smart Transportation” and “Digital Entrepreneurship”. Let’s talk and discuss about the rationale for these new emerging areas/topics.

- “Smart Transportation” (Subset of Vocabulary: transport, route, sensor)** - Since the beginning of civilization, transportation has been a necessity for humanity. As technology developed, information and communications technology (ICT) was added to intelligent transportation systems. The Internet of Things (IoT) and ICT are being used to create "smart cities," which are already being used as a tool to increase the effectiveness of city operations and services. Recently, a number of IoT-based smart city apps have been created. Among these uses, smart transportation services are essential for tackling issues like traffic control, parking for cars, and road safety (Fantin Irudaya Raj & Appadurai, 2022). Since the transportation system is normally constructed to accommodate routine demand, disruptions brought on by exceptional events are a well-known concern in the transportation industry. The inability to gather thorough and trustworthy data early enough to prepare mitigation actions is one aspect of the issue. Many cities across the world would benefit greatly from a program that automatically searches the internet for events and forecasts their effects on

transportation planning. Markou et al. (2019) focuses on the difficulties in finding and interpreting web content regarding actual occurrences in order to use it for demand explanation and forecast. A crucial area of research for traffic safety planning is the extraction of traffic features. Road network features had previously been manually retrieved in traffic tasks. Aiming to automatically learn low-dimensional node representations, network representation learning seeks to contrast this. Representation learning of urban nodes is investigated as a supervised task in light of feature learning in NLP (Huang et al., 2020). To develop feature representations for road network nodes based on trip paths, a deep learning framework is also suggested. The identification and condition analysis of traffic accidents utilizing ontology LDA and bidirectional LSTM are proposed in a framework for real-time monitoring via social networks (Ali et al., 2021).

- **“Digital Entrepreneurship” (Subset of Vocabulary: digital entrepreneurship, collaboration, ceo)** - COVID-19 has urged numerous organizations to concentrate on implementing digital technology in order to survive. As a result of the expanding digital wave during COVID-19, there have been several prospects for new firms to enter the market (Modgil et al., 2022). The COVID-19 epidemic has pushed this move towards increased automation as digital ecosystems accept the needs of digital workers and bots. Wilk et al. (2021) looked at social media activity in terms of significant themes, patterns, and online mood about digital entrepreneurship. It offers a moment-in-time, visually-focused view on user-generated content (UGC) on social media to help people better comprehend the subject of digital business. The knowledge-based economy relies heavily on university-industry collaborations (UICs), De Silva et al. (2021) finds that using sentiment analysis on a dataset of 415 final reports from completed UICs, it was discovered that there is a negative correlation between the collaborators' reported issues and benefits of UICs, which is mediated by negative emotional assessment. Developing media communications that effectively portray the company to the outside world is a crucial task for new venture CEOs. Performance is significantly impacted by these choices and actions. Howard et al. (2021) investigates the interaction between founder CEOs and firm media strategy in their influence on initial public offering (IPO) with using textual analysis and NLP. According to the findings, media coverage that is more frequent and uses more positive language mediates the impacts of founding CEOs favorably, increasing the likelihood of a



like personal computers, tablets, and even smartphones have improved teaching methods and transformed the educational system. Text analytics assists teachers in evaluating student performance, determining how similar a lecturer's and students' posts are in a discussion forum, and gathering student feedback on a teaching method in order to categorize each student's feedback. Mohammed et al. (2021) emphasize the key elements of IoT analytics and provide a thorough foundation of the methods and applications of text analytics. Among other NLP tasks, emotion recognition has profited considerably from the usage of massive transformer models. However, because these models require a lot of computation, deploying them on devices with limited resources is extremely difficult. Using massive transformers as high-quality feature extractors and straightforward classifiers based on linear separators that are friendly to hardware, Pandealea et al. (2021) demonstrate in this work that it is possible to attain superior performance while enabling real-time inference and rapid training. Multiple running rules that act on actuators in conflicting ways may result in unanticipated and unpredictable interference issues with smart home systems, which could endanger occupants and their possessions. Xiao et al. (2019) suggest an automatic interference detection technique based on knowledge graphs. utilizing lexical databases and natural NLP techniques.

- **“Business Management” (Subset of Vocabulary: productivity, business process, industry)** - With the rapid advancement of the academic discipline, increased digitization has enabled innovation and practical examples from the several business areas that have attracted the attention of marketing researchers. During the COVID-19 crisis, the rate of change has dramatically accelerated. Since the entertainment sector depends so largely on producing engaging material for consumers, increasing productivity is a tremendously difficult undertaking. The consumer-centric design approach, which prioritizes the needs of the consumer in the creation of all content, aims to help companies create services and goods that are specifically catered to the needs of their target market. Del Vecchio et al. (2021) provide a novel framework with using NLP and econometric that enables the application of data science to optimize content creation in the entertainment sector and test this framework for the film sector. The monitoring service is a crucial component of business process management that can stop a number of issues before they arise in businesses and industries. A computer system with knowledge of the business process creates an execution log, which aids in



process prediction. Moon et al. (2021) aim to forecast events based on completed event log data and estimate the process that will run after the currently running process instance. Critical contemporary concerns include multi-tier supply chains in Industry 4.0. Zhou et al. (2021) briefly analyses the Industry 4.0 regulations in several nations. A comparison of the coherence values for two LDA-based machine learning classification algorithms was done in order to determine which model to use and how many topics to include in the model.

For C4 – Machine Learning Models/Types, “**bilstm-cnn**” word appears in the cluster. “bilstm-cnn” model can be used/utilized by NER and BERT models/algorithms separately or jointly. As these application domain related words also appear in 2027 December prediction, it can be expected to sustain the popularity of “bilstm-cnn” word/method.

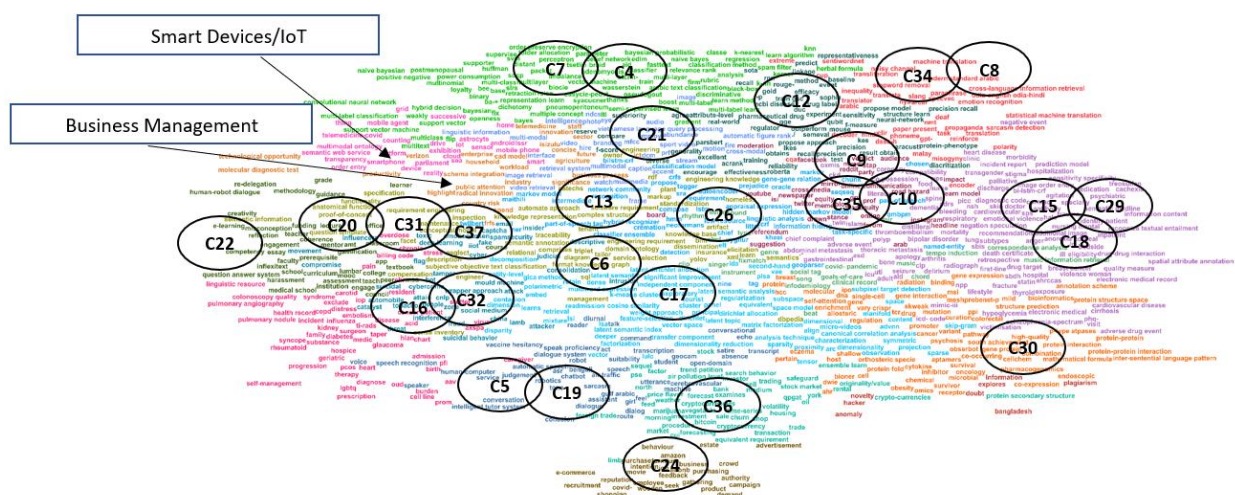


Figure 54. 2030 December Predict

## Chapter 4

### Conclusions

#### 4.1 Discussions

In this section, discussions are divided in two parts. In the first part, results are evaluated for opinion mining use case. Second part focuses on opinion forecasting use case.

**4.1.1 Opinion mining.** As illustrated in Table 1, sentiment analysis and NLP approaches have been proven to be effective techniques for product or technology reviews and for retrieving opinion from the relevant community. These approaches are even more effective when the relevant datasets are difficult to examine due to time, budget or resource constraints. Our practical and methodological findings show that the proposed method in this study is an effective one considering the performance metrics and the specific case of Google Glass. Following studies that aim to utilize social media data to retrieve novel ideas for product development (Li et al., 2014; Mirtalaie et al., 2017), we extend the current knowledge and relevant applications considering the proposed model – “social media-based opinion retrieval using multi-task deep neural networks” – as well as the application area of the model (product analysis), in order to identify new use cases for an innovative product and support product development-oriented innovation activities.

The interpretations of the results can be further extended by focusing on both academic and methodological aspects. Starting from the academic perspective, previous literature has focused on product analysis with different approaches; Lipizzi (2015) and Nuortimo and Harkonen (2018) examined it from a product launch perspective, Rane and Kumar (2018) and Ibrahim and Wang (2019) examined negative feedback, and some authors have examined sentiments for companies (Botchway et al., 2019) or products (Basiri et al., 2020). In terms of new product development and innovations, many of these are appropriate for incremental innovations where products are improved upon the previous offering but may not be highly suitable for radical innovations (Holahan et al., 2014). However, relative to previous approaches, our results show that our model is highly suitable for both incremental and radical innovations as we identify radical (game changing) ideas as well as incremental ones (improvements based on the previous offerings).

To extend further the discussion on the suitability of our model for new product development and innovation practices, we should consider the criticism of Trott (2001) that consumer opinion is not suitable, especially for radical innovations, particularly where there is asymmetric knowledge between the companies and consumers. In our case, we found that when knowledge is being retrieved from a large database with a systematic approach, it is possible to organize the relevant information to be beneficial, even for high-tech products. Furthermore, it is possible to tap into the hidden information or market positioning opportunities (Google Glass may be better suited for industrial or scientific actors instead of direct consumers). For example, Mirtalaie et al. (2017) successfully retrieved product feature-specific information and performed cross-domain analysis to identify novel ideas that can be integrated into future generations of the reference product (See Table 1). We extend this information by uncovering new product application areas, product improvement opportunities and next generation innovations.

We also make methodological contributions to the related literature. As illustrated in Table 1, a detailed product analysis from unstructured data consisting of user comments requires the design of a multi-stage framework. The proposed opinion retrieval system in this study is also based on two machine learning tasks which address sentiment analysis and opinion detection problems. We used the opinion detection module to identify the user comments that include useful feedback or suggestions about the related product. Thus, we removed the comments which did not include useful feedback but only expressed sentiment regarding the reference product. Combining the outputs of the models allowed us to conduct a detailed sentiment-based analysis of the user opinions with a word network map consisting of keyword clusters appearing in the related tweets. By training these tasks with a multi-task DNN architecture, we improved the generalization ability of the resulting system.

Finally, we should also mention transferability and further application areas of our model. Although we applied it to the consumer electronics area, others can implement it in other areas by training the model with specific data and labelling steps. The proposed model could also be implemented for service innovation or business model innovation activities.

**4.1.2 Technological forecasting.** As illustrated in Table 17, word embedding estimation with LSTM method has been proven to be effective technique for technology prediction and

emerging topic/cluster analysis. Our practical and methodological findings show that the proposed method in this study is highly effective considering the cluster similarity metrics and also in the specific case of Text Mining.

Comparing our approach to the previously published technology foresight and forecasting methods, we can say that our method has both advantages and disadvantages. Our method is useful to identify possible changes and emerging areas in the near future. However, it may not be an accurate estimation of future. However, it may be a more accurate method compared to other emergence related calculations and methods to identify the changing landscape. Previously published studies identified the emergence of technologies once it was emerged in the field. However, our model can identify the emergence of the fields by predicting future trends.

We also make methodological contributions to the related literature. As illustrated in Table 2, the dominant method is topic modelling techniques/algorithms for both Group 1: Trend & Topic Analysis in Science, Technology and Innovation and Group 2: Technology Forecasting & Foresighting domains. The proposed forecasting system in this study introduces applying deep learning technique, LSTM model per word individually based on identified vocabulary. In addition to, rather than predicting next month or year, the situation/case after next 3 years is aimed to be predicted. Also, if we compare our model to the statistical and trend predictions, in our results, we do not only predict the new areas but their relationship compared to the clusters that are already there in the actual results. Our model recognises the technological interdependence and their relationship in the future. This logic and how it improved the technological forecasting is highlighted in our results considering the clustering visuals and Table 17. Finally, we can mention the weakness of our model. Although our model improved the technology prediction models with a matrix based approach, our model could not identify the new terms or technologies if they have never been used in the literature. However, this issue can be resolved with a comprehensive foresight model where the predictions are also supported by multiple datasets such as patent data and also expert opinion.

## **4.2 Contributions**

In this section, contributions are divided in two parts. In the first part, contributions are evaluated for opinion mining use case. Second part focuses on opinion forecasting use



case

**4.2.1 Opinion mining.** In this study, we proposed a framework that uses social media data to reveal the reasons for a failed innovative product from the customer perspective and to suggest new use cases and innovative ideas for product development. To do so, we applied various single and multi-task DNN models with different word representation techniques to perform opinion retrieval based on social media. As a case study, Google Glass was selected using the Twitter platform. For the main opinion retrieval task, we identified two supervised learning tasks as sentiment prediction and opinion detection. Sentiment prediction was handled as a multi-class problem with negative, neutral and positive class labels representing the sentiment of the user's opinion about the related product. We designed the opinion detection module as a binary-class problem representing the existence of a product-related opinion in the related content.

According to the findings of our experiments, multi-task DNN models considerably outperformed single-task DNN models in F1 scores for both sentiment prediction and opinion detection tasks. From the point of feature input set, the best results were obtained using Word2Vec features individually and Word2Vec and GloVe features together. The best performing model, which yielded the highest F1 scores for both tasks, was the multi-task DNN architecture with Word2Vec features. We applied this model to the unlabeled tweets to obtain the predictions for both target variables and used the predicted labels to eliminate the tweets that do not contain any feedback or suggestion. We then identified and analyzed the remaining negative tweets and positive + neutral tweets separately, then visualized the obtained results as a word network map of the related terms.

This work adds to the literature in both practical and methodological ways. The main practical contribution is the resulting opinion retrieval framework, which may be used as a decision support system for product development approaches. It can be integrated into different phases of an NPD process for both incremental and radical innovations. The implementation of this approach to incremental innovations could make it easier to identify problems that can be separated in the next generation of products. Radical innovation may present more difficulties as it would require a complete redesign of products considering the feedback received at a holistic level. For the case of Google Glass, the findings suggest a redesign and radical innovation are required based on both the negative and positive opinions retrieved.

The main methodological contribution of this study is the design of a multi-task DNN architecture that learns two supervised learning tasks simultaneously which are required for a sentiment-based detailed opinion retrieval system. The results confirm that a multi-task DNN model offers better generalization ability for both sentiment prediction and opinion detection tasks when compared to two single-task DNN models trained independently for each task. Another contribution is the proposal of an end-to-end system for product-related sentiment-based opinion retrieval from social media. The system, from data collection to the sentiment-based visualization of the keyword maps, offers a detailed product analysis framework for decision makers.

Our study contributes to the previous literature as shown below:

- Existing studies mostly analyze only the tweets labelled as negative by the sentiment analysis module to present the products and services about which the customers complain most. Our proposed system uses the opinion detection module to also detect and analyze the positive and neutral comments that may contain useful feedback or suggestion that can be utilized for product development, innovation-oriented decisions and finding new use cases for the related product,
- Existing studies mostly use domain-specific features or lexicons for data representation, whereas we propose a generalizable model using state-of-the-art word embedding techniques that can be applied to any social media-based text data,
- Most of the existing studies apply multiple phases for social-media analysis. Our findings show that the overall accuracy of the system can be improved by training these tasks simultaneously with a multi-task learning approach,
- Most of the existing studies present some statistical results, such as the frequencies of sentiment labels of the user reviews about the related products or services, whereas the proposed system in this study presents word network maps to summarize the sentiment-based opinions that can be directly utilized for product development or to determine new use-cases,
- Our proposed system clusters and identifies customer opinion considering the sentimental classifications specific to the product development and innovation

opportunities.

Our results have implications for relevant practitioners such as product development specialists and also have implications specific to the smart glass industry. Our proposed model calls for the relevant practitioners to implement smart or advanced approaches into product development practices to minimize failure rates in commercialization activities. Our selected case highlights how technological advancements alone are not adequate for successful innovations. In the case of Google Glass, it was apparent that other factors, such as legality, safety and usability aspects, were some of the key reasons why such a great product had lower adoption or diffusion rates. Specific to the smart glass industry, relevant companies need to work on an enhanced design for the general public that considers design, hardware and legal issues. However, the public's privacy concerns may be the most difficult challenge to overcome the results illustrate that smart glass development strategies appear to be more suitable for specific markets (i.e., workplace and productivity use cases). Considering industry-specific application areas of smart glasses, the involvement of key stakeholders may be the most crucial aspect for technological acceptance considering the required software applications as well as industry specific know-how. To resolve this issue, companies such as Google may need to collaborate with industry-specific partners at a global scale to accelerate the diffusion of this technology.

**4.2.2 Technological forecasting.** In this study, we proposed a framework that uses literature data to learn the existing topics and clusters for selected technology, use-case and to predict new clusters, topics for the future and identify emerging areas. To do so, we applied various techniques to be able to construct/identify the vocabulary evolution month by month. We treated trained word embedding matrixes regarding the vocabulary as a time-series data and used as an input to LSTM network to predict the view of next 3 years snapshot/observation. This approach was the first one in TF literature.

As a case study, Text Mining sub-domain under Computer Science and AI was selected using the WoS platform. We identified LSTM network as supervised learning regression task. For the visualization part, we used cosine similarity of word pairs for both actual and predicted values. We generated clusters based on the calculated similarity values. The results are also interesting and useful for the text mining/intelligence community to see

the changing landscape in this field.

Our experimental results show that proposed LSTM model performed well from the point of clustering evaluation regarding 2021 December Actual and Predict clusters. In addition to, the proposed model is able to catch all the changed/evolved clusters from 2018 December to 2021 December. For evaluation purposes, we also proposed a measurement model in Table 17. As such method was never been implemented before, we introduced a measurement model for other scholars to implement in their studies to assess their results' performance.

This research adds methodological and practical elements to the literature. The main practical contribution is the resulting technology forecasting framework, which can be used as a technology foresighting system for decision-makers, strategic planners, technology enthusiasts and innovators. It can be integrated into different departments of research and development companies, institutes and industry.

The main methodological contribution of this study is the design of a LSTM architecture that learns past embedding features of a word as time-series data and predict future outlook of word embedding. Another contribution is the proposal of an end-to-end system for technology forecasting based word embedding matrix trained from literature data.

Our study contributes to the previous literature as shown below:

- Existing studies mostly focus on traditional unsupervised topic modelling algorithms. Our proposed system combines semantic word embedding features with deep learning method to be treated as time-series data.
- Existing studies mostly uses feature in document level. Our proposed system applies LSTM model in word level.
- Most of the existing studies do not consider evolution of clusters in a predicted format. Our study tries to compare changes in the cluster appearance and also predicts evolution of clusters for a future timeline.
- Our proposed system is capable to predict the future embedding outlook of a new word with using the most similar word in the training data.

- A new performance evaluation metrics are proposed for other researchers to implement and evaluate their work

### 4.3 Limitations and Future Works

In this section, limitations and future works is divided in two parts. In the first part, limitations and future works are evaluated for opinion mining use case. Second part focuses on opinion forecasting use case.

**4.3.1 Opinion mining.** Future research can further enhance the proposed opinion retrieval system using bidirectional transformers for word representation such as BERT and XLNet. In addition, unlabeled tweets can be utilized during training by a semi-supervised learning approach with the aim of improving the accuracy of the overall system and also reducing the required number of labelled samples for a robust model. Furthermore, the proposed model can be tested in different areas such as low-tech environments to analyze its generalizability and performance. Other researchers can implement our model for service innovations or business model innovations. Finally, further opinion retrieval systems can be designed for different phases of the product development process.

**4.3.2 Technological forecasting.** Based on literature review, it is found that there does not exist a similar study with approaching and applying technological forecasting in that way. Further studies on that domain can use/benefit this proposed model and results as a checkpoint to compare their results. Future research can further enhance the proposed technological forecasting system using bidirectional transformers for word representation such as BERT, XLNet and ElMo. In addition, unsupervised topic modelling techniques such as LDA, SVD could be applied for generating topics/clusters as an alternative. Topic modelling methods can also be used for reducing dimensionality of vocabulary before creating cosine similarity matrix. Different time step values can be also utilized inside LSTM network. Furthermore, the proposed model can be tested in different areas such Blockchain to identify emerging areas or in interdisciplinary areas such as Biochemistry to see changing landscape how different fields possibly merge together in the future.

## REFERENCES

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2019). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv 2016. arXiv preprint arXiv:1603.04467.
- Aggarwal, C. C. (2015). *Data mining: the textbook (Vol. 1)*. New York: springer.
- Aggarwal, C. C. (2018). *Neural networks and deep learning*. Springer, 10, 978-3.
- Ali, F., Ali, A., Imran, M., Naqvi, R. A., Siddiqi, M. H., & Kwak, K. S. (2021). Traffic accident detection and condition analysis based on social networking data. *Accident Analysis & Prevention*, 151, 105973.
- Almeida, F., & Xexéo, G. (2019). Word embeddings: A survey. arXiv preprint arXiv:1901.09069.
- Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.
- Alshemali, B., & Kalita, J. (2020). Improving the reliability of deep neural networks in NLP: A review. *Knowledge-Based Systems*, 191, 105210.
- Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *R&D Management*, 50(3), 329-351.
- Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77, 236-246.
- Baroni, M., Dinu, G., & Kruszewski, G. (2014, June). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 238-247).

- Baror, S. O., Venter, H. S., & Adeyemi, R. (2021). A natural human language framework for digital forensic readiness in the public cloud. *Australian Journal of Forensic Sciences*, 53(5), 566-591.
- Basiri, M. E., Abdar, M., Cifci, M. A., Nemati, S., & Acharya, U. R. (2020). A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques. *Knowledge-Based Systems*, 198, 105949.
- Bastian, M., Heymann, S., & Jacomy, M. (2009, March). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media (Vol. 3, No. 1, pp. 361-362)*.
- Batagelj, V., & Mrvar, A. (2004). Pajek—analysis and visualization of large networks. In *Graph drawing software (pp. 77-103)*. Springer, Berlin, Heidelberg.
- Behpour, S., Mohammadi, M., Albert, M. V., Alam, Z. S., Wang, L., & Xiao, T. (2021). Automatic trend detection: Time-biased document clustering. *Knowledge-Based Systems*, 220, 106907.
- Bengfort, B., Danielsen, N., Bilbro, R., Gray, L., McIntyre, K., Richardson, G., ... & Keung, J. (2018). Yellowbrick. URL <http://www.scikit-yb.org/en/latest>.
- Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13.
- Bengisu, M., & Nekhili, R. (2006). Forecasting emerging technologies with the aid of science and technology databases. *Technological Forecasting and Social Change*, 73(7), 835-844.
- Berger, A., Vokalova, A., Maly, F., & Poulova, P. (2017, August). Google glass used as assistive technology its utilization for blind and visually impaired people. In *International Conference on Mobile Web and Information Systems (pp. 70-82)*. Springer, Cham.

- Bhoir, S., Ghorpade, T., & Mane, V. (2017, December). Comparative analysis of different word embedding models. In *2017 International conference on advances in computing, communication and Control (ICAC3)* (pp. 1-4). IEEE.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."
- Bisong, E. (2019). *Building machine learning and deep learning models on Google cloud platform: A comprehensive guide for beginners*. Apress.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5, 135-146.
- Botchway, R. K., Jibril, A. B., Kwarteng, M. A., Chovancova, M., & Oplatková, Z. K. (2019, September). A review of social media posts from UniCredit bank in Europe: A sentiment analysis approach. In *Proceedings of the 3rd international conference on business and information Management* (pp. 74-79).
- Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30, 31-40.
- Bukhari, A. H., Raja, M. A. Z., Sulaiman, M., Islam, S., Shoaib, M., & Kumam, P. (2020). Fractional neuro-sequential ARFIMA-LSTM for financial market forecasting. *IEEE Access*, 8, 71326-71338.
- Burmaoglu, S., Sartenaer, O., Porter, A., & Li, M. (2019). Analysing the theoretical roots of technology emergence: an evolutionary perspective. *Scientometrics*, 119(1), 97-118.
- Cabezas, J., Moctezuma, D., Fernández-Isabel, A., & Martín de Diego, I. (2021). Detecting emotional evolution on twitter during the covid-19 pandemic using text analysis. *International Journal of Environmental Research and Public Health*, 18(13), 6981.
- Carrera, J. F., Wang, C. C., Clark, W., & Southerland, A. M. (2019). A systematic review



- of the use of google glass in graduate medical education. *Journal of Graduate Medical Education*, 11(6), 637-648.
- Chandrayan, S., & Bamne, P. (2021). A brief survey of Text Mining and its applications. *International Journal*, 9(8).
- Chang, W., & Taylor, S. A. (2016). The effectiveness of customer participation in new product development: A meta-analysis. *Journal of Marketing*, 80(1), 47-64.
- Chen, H., Zhang, G., Zhu, D., & Lu, J. (2017). Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014. *Technological Forecasting and Social Change*, 119, 39-52.
- Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.
- Chollet, F., & others. (2015). Keras. GitHub. Retrieved from <https://github.com/fchollet/keras>
- Chowdhary, K. (2020). *Natural language processing. Fundamentals of artificial intelligence*, 603-649.
- Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167).
- Cummins, N., Amiriparian, S., Ottl, S., Gerczuk, M., Schmitt, M., & Schuller, B. (2018, April). Multimodal bag-of-words for cross domains sentiment analysis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4954-4958). IEEE.
- Dang, S., & Ahmad, P. H. (2015). A review of text mining techniques associated with various application areas. *International Journal of Science and Research (IJSR)*, 4(2), 2461-2466.

- Datta, P., Lodinger, N., Namin, A. S., & Jones, K. S. (2020, December). Predicting consequences of cyber-attacks. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 2073-2078). IEEE.
- De Silva, M., Rossi, F., Yip, N. K., & Rosli, A. (2021). Does affective evaluation matter for the success of university-industry collaborations? A sentiment analysis of university-industry collaborative project reports. *Technological Forecasting and Social Change*, *163*, 120473.
- Del Vecchio, M., Kharlamov, A., Parry, G., & Pogrebna, G. (2021). Improving productivity in Hollywood with data science: Using emotional arcs of movies to drive product and service innovation in entertainment industries. *Journal of the Operational Research Society*, *72*(5), 1110-1137.
- Delgado López-Cózar, E., Orduña-Malea, E., & Martín-Martín, A. (2019). Google Scholar as a data source for research assessment. In *Springer handbook of science and technology indicators* (pp. 95-127). Springer, Cham.
- Delgosha, M. S., Hajiheydari, N., & Talafidaryani, M. (2021). Discovering IoT implications in business and management: a computational thematic analysis. *Technovation*, 102236.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.
- Denter, N. M., Aaldering, L. J., & Caferoglu, H. (2022). Forecasting future bigrams and promising patents: introducing text-based link prediction. *foresight*, (ahead-of-print).
- Dhillon, P., Foster, D. P., & Ungar, L. (2011). Multi-view learning of word embeddings via cca. *Advances in neural information processing systems*, *24*.
- Ding, Y., Zhu, Y., Feng, J., Zhang, P., & Cheng, Z. (2020). Interpretable spatio-temporal

- attention LSTM model for flood forecasting. *Neurocomputing*, 403, 348-359.
- Djelassi, S., & Decoopman, I. (2013). Customers' participation in product development through crowdsourcing: Issues and implications. *Industrial Marketing Management*, 42(5), 683-692.
- Du, J., Preston, S., Sun, H., Shegog, R., Cunningham, R., Boom, J., ... & Tao, C. (2021). Using Machine Learning–Based Approaches for the Detection and Classification of Human Papillomavirus Vaccine Misinformation: Infodemiology Study of Reddit Discussions. *Journal of medical Internet research*, 23(8), e26478.
- Ergen, T., & Kozat, S. S. (2017). Online training of LSTM networks in distributed systems for variable length data sequences. *IEEE transactions on neural networks and learning systems*, 29(10), 5159-5165.
- Escamilla-García, A., Soto-Zarazúa, G. M., Toledano-Ayala, M., Rivas-Araiza, E., & Gastélum-Barrios, A. (2020). Applications of artificial neural networks in greenhouse technology and overview for smart agriculture development. *Applied Sciences*, 10(11), 3835.
- Fantin Irudaya Raj, E., & Appadurai, M. (2022). Internet of Things-Based Smart Transportation System for Smart Cities. In *Intelligent Systems for Social Good* (pp. 39-50). Springer, Singapore.
- Filippov, S., & Hofheinz, P. (2016). Text and Data Mining for Research and Innovation. *Learned Publishing*, 23, 3.
- Gaur, V., & Kumar, R. (2022). Analysis of Machine Learning Classifiers for Early Detection of DDoS Attacks on IoT Devices. *Arabian Journal for Science and Engineering*, 47(2), 1353-1374.
- Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1), 1-309.

- Goyal, P., Pandey, S., & Jain, K. (2018). *Deep learning for natural language processing*. New York: Apress.
- Gundecha, U. (2014). *Learning Selenium testing tools with Python*. Packt Publishing Ltd.
- Gupta, R. K., Agarwalla, R., Naik, B. H., Evuri, J. R., Thapa, A., & Singh, T. D. (2022). Prediction of Research Trends using LDA based Topic Modeling. *Global Transitions Proceedings*.
- Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, 1(1), 60-76.
- Gusenbauer, M. (2019). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1), 177-214.
- HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PloS one*, 15(5), e0232525.
- Hameed, I., Khan, M. B., Shahab, A., Hameed, I., & Qadeer, F. (2016). Science, technology and innovation through entrepreneurship education in the United Arab Emirates (UAE). *Sustainability*, 8(12), 1280.
- Hao, T., Chen, X., Li, G., & Yan, J. (2018). A bibliometric analysis of text mining in medical research. *Soft Computing*, 22(23), 7875-7892.
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Berg, S. (2020). Smith 474 nj. Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del R'io JF, Wiebe M, Peterson P, G'erard-475 Marchant P, et al. Array programming with NumPy. *Nature*, 585(7825), 357-362.
- Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing*, 4(1), 1.

- Hasson, S. G., Piorkowski, J., & McCulloh, I. (2019, August). Social media as a main source of customer feedback: alternative to customer satisfaction surveys. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 829-832).
- Heo, S., & Lee, J. H. (2018). Fault detection and classification using artificial neural networks. *IFAC-PapersOnLine*, 51(18), 470-475.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
- Holahan, P. J., Sullivan, Z. Z., & Markham, S. K. (2014). Product development as core competence: How formal product development practices differ for radical, more innovative, and incremental product innovations. *Journal of Product Innovation Management*, 31(2), 329-345.
- Hou, T., Yannou, B., Leroy, Y., & Poirson, E. (2019). Mining customer product reviews for product development: A summarization process. *Expert Systems with Applications*, 132, 141-150.
- Howard, M. D., Kolb, J., & Sy, V. A. (2021). Entrepreneurial identity and strategic disclosure: Founder CEOs and new venture media strategy. *Strategic Entrepreneurship Journal*, 15(1), 3-27.
- Hu, K., Luo, Q., Qi, K., Yang, S., Mao, J., Fu, X., ... & Zhu, Q. (2019). Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis. *Information Processing & Management*, 56(4), 1185-1203.
- Huang, L., Liu, F., & Zhang, Y. (2020). Overlapping Community Discovery for Identifying Key Research Themes. *IEEE Transactions on Engineering Management*, 68(5), 1321-1333

- Huang, S., Shao, C., Li, J., Yang, X., Zhang, X., Qian, J., & Wang, S. (2020). Feature Extraction and Representation of Urban Road Networks Based on Travel Routes. *Sustainability*, 12(22), 9621.
- Huang, H. C., & Su, H. N. (2019). The innovative fulcrums of technological interdisciplinarity: An analysis of technology fields in patents. *Technovation*, 84, 59-70.
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1), 193-218.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(03), 90-95.
- Hussain, S., Muhammad, L. J., & Yakubu, A. (2018). Mining social media and DBpedia data using Gephi and R. *Journal of Applied Computer Science & Mathematics*, 12(1), 14-20.
- Ibrahim, N. F., & Wang, X. (2019). A text analytics approach for online retailing service improvement: Evidence from Twitter. *Decision Support Systems*, 121, 37-50.
- Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). PMLR.
- Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one*, 9(6), e98679.
- Jatnika, D., Bijaksana, M. A., & Suryani, A. A. (2019). Word2vec model analysis for semantic similarities in english words. *Procedia Computer Science*, 157, 160-167.
- Jebari, C., Herrera-Viedma, E., & Cobo, M. J. (2021). The use of citation context to detect

- the evolution of research topics: a large-scale analysis. *Scientometrics*, 126(4), 2971-2989.
- Jee, J., Shin, H., Kim, C., & Lee, S. (2021). Six different approaches to defining and identifying promising technology through patent analysis. *Technology Analysis & Strategic Management*, 1-13.
- Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, 24(10), 2733-2742.
- Jia, Y., Hoberock, J., Garland, M., & Hart, J. (2008). On the visualization of social and other scale-free networks. *IEEE transactions on visualization and computer graphics*, 14(6), 1285-1292.
- Jiang, H., Kwong, C. K., Kremer, G. O., & Park, W. Y. (2019). Dynamic modelling of customer preferences for product design using DENFIS and opinion mining. *Advanced Engineering Informatics*, 42, 100969.
- Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011, June). Target-dependent twitter sentiment classification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies* (pp. 151-160).
- Jin, G., & Yu, Z. (2021). A Korean named entity recognition method using Bi-LSTM-CRF and masked self-attention. *Computer Speech & Language*, 65, 101134.
- Jissink, T., Schweitzer, F., & Rohrbeck, R. (2019). Forward-looking search during innovation projects: Under which conditions it impacts innovativeness. *Technovation*, 84, 71-85.
- Jung, H., & Lee, B. G. (2020). Research trends in text mining: Semantic network and main path analysis of selected journals. *Expert Systems with Applications*, 162, 113851.

- Kanani, P., & Padole, M. (2019). Deep learning to detect skin cancer using google colab. *International Journal of Engineering and Advanced Technology Regular Issue*, 8(6), 2176-2183.
- Kayte. (2021, November 11) Problems with Classification Examples from Real Life <https://medium.datadriveninvestor.com/problems-with-classification-examples-from-real-life-645b7b756e96>
- Khan, M., Khan, M. S., & Alharbi, Y. (2020). Text Mining Challenges and Applications, A Comprehensive Review. *IJCSNS*, 20(12), 138.
- Kim, G., & Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. *Technological Forecasting and Social Change*, 117, 228-237.
- Kim, J., & Geum, Y. (2021). How to develop data-driven technology roadmaps: The integration of topic modeling and link prediction. *Technological Forecasting and Social Change*, 171, 120972.
- Kim, S., Park, H., & Lee, J. (2020). Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications*, 152, 113401.
- Kostoff, R. N., & Schaller, R. R. (2001). Science and technology roadmaps. *IEEE Transactions on engineering management*, 48(2), 132-143.
- Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights*, 1(1), 100008.
- Lannelongue, L., Grealey, J., & Inouye, M. (2021). Green algorithms: quantifying the carbon footprint of computation. *Advanced science*, 8(12), 2100707.
- Lebret, R., & Collobert, R. (2013). Word emdeddings through hellinger pca. arXiv preprint



arXiv:1312.5542.

- Lee, S., Choi, J., & Sawng, Y. W. (2019). Foresight of promising technologies for healthcare-IoT convergence service by patent analysis.
- Li, Y. M., Chen, H. M., Liou, J. H., & Lin, L. F. (2014). Creating social intelligence for product portfolio design. *Decision Support Systems*, 66, 123-134.
- Li, B., Lin, Y., & Zhang, S. (2017). Multi-task learning for intrusion detection on web logs. *Journal of Systems Architecture*, 81, 92-100.
- Li, Q., & Shah, S. (2017, August). Learning stock market sentiment lexicon and sentiment-oriented word vector from stocktwits. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (pp. 301-310).
- Li, Y., & Yang, T. (2018). Word embedding for understanding natural language: a survey. In *Guide to big data applications* (pp. 83-104). Springer, Cham.
- Li, X., Xie, Q., Daim, T., & Huang, L. (2019). Forecasting technology trends using text mining of the gaps between science and technology: The case of perovskite solar cell technology. *Technological Forecasting and Social Change*, 146, 432-449.
- Li, S., Zhu, J., & Miao, C. (2015). A generative word embedding model and its low rank positive semidefinite solution. arXiv preprint arXiv:1508.03826.
- Liang, Z., Mao, J., Lu, K., Ba, Z., & Li, G. (2021). Combining deep neural network and bibliometric indicator for emerging research topic prediction. *Information Processing & Management*, 58(5), 102611.
- Lipizzi, C., Iandoli, L., & Marquez, J. E. R. (2015). Extracting and evaluating conversational patterns in social media: A socio-semantic analysis of customers' reactions to the launch of new products using Twitter streams. *International Journal of Information Management*, 35(4), 490-503.

- Liu, Y., Jiang, C., & Zhao, H. (2019). Assessing product competitive advantages from the perspective of customers by mining user-generated content on social media. *Decision Support Systems*, 123, 113079.
- Liu, J., Wei, J., & Liu, Y. (2021). Technology Forecasting based on Topic Analysis and Social Network Analysis: A Case Study Focusing on Gene Editing Patents. *Journal of Scientific and Industrial Research (JSIR)*, 80(05), 428-437.
- Loria, S. (2018). textblob Documentation. Release 0.15, 2.
- Loveridge, D. (2008). *Foresight: The art and science of anticipating the future*. Routledge.
- Lu, W., Huang, S., Yang, J., Bu, Y., Cheng, Q., & Huang, Y. (2021). Detecting research topic trends by author-defined keyword frequency. *Information Processing & Management*, 58(4), 102594.
- Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2), 203-208.
- Lyu, J. C., Le Han, E., & Luli, G. K. (2021). COVID-19 vaccine-related discussion on Twitter: topic modeling and sentiment analysis. *Journal of medical Internet research*, 23(6), e24435.
- Mai, L., & Le, B. (2021). Joint sentence and aspect-level sentiment analysis of product comments. *Annals of Operations research*, 300(2), 493-513.
- Markou, I., Kaiser, K., & Pereira, F. C. (2019). Predicting taxi demand hotspots using automated Internet Search Queries. *Transportation Research Part C: Emerging Technologies*, 102, 73-86.
- Martelli, A., Ravenscroft, A. M., & Holden, S. (2017). *Python in a Nutshell*. O'Reilly Media.

- Massey, D., Huang, C., Lu, Y., Cohen, A., Oren, Y., Moed, T., ... & Krumholz, H. (2021). Engagement with COVID-19 public health measures in the United States: A cross-sectional social media analysis from June to November 2020. *Journal of medical Internet research*, 23(6), e26655.
- McKinney, W. (2010, June). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference (Vol. 445, No. 1, pp. 51-56)*.
- Meissner, D., & Sokolov, A. (2013). Foresight and science, technology and innovation indicators. In *Handbook of innovation indicators and measurement*. Edward Elgar Publishing.
- Meng, Y., Huang, J., Wang, G., Zhang, C., Zhuang, H., Kaplan, L., & Han, J. (2019). Spherical text embedding. *Advances in Neural Information Processing Systems*, 32.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- Miles, I., Saritas, O., & Sokolov, A. (2016). *Foresight for science, technology and innovation*. Switzerland: Springer International Publishing.
- Mirtalaie, M. A., Hussain, O. K., Chang, E., & Hussain, F. K. (2017). A decision support framework for identifying novel ideas in new product development from cross-domain analysis. *Information Systems*, 69, 59-80.
- Mnih, A., & Hinton, G. (2007, June). Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning* (pp. 641-648).
- Modgil, S., Dwivedi, Y. K., Rana, N. P., Gupta, S., & Kamble, S. (2022). Has Covid-19 accelerated opportunities for digital entrepreneurship? An Indian perspective. *Technological Forecasting and Social Change*, 175, 121415.

- Mohamed, T., Sayed, S., Salah, A., & Houssein, E. H. (2021). Long short-term memory neural networks for RNA viruses mutations prediction. *Mathematical Problems in Engineering*, 2021.
- Mohammed, A. H. K., Jebamikyous, H. H., Nawara, D., & Kashef, R. (2021). IoT text analytics in smart education and beyond. *Journal of Computing in Higher Education*, 33(3), 779-806.
- Mohd, M., Jan, R., & Shah, M. (2020). Text document summarization using word embedding. *Expert Systems with Applications*, 143, 112958.
- Moon, J., Park, G., & Jeong, J. (2021). Pop-on: Prediction of process using one-way language model based on nlp approach. *Applied Sciences*, 11(2), 864.
- Morin, F., & Bengio, Y. (2005, January). Hierarchical probabilistic neural network language model. In *International workshop on artificial intelligence and statistics* (pp. 246-252). PMLR.
- Mun, C., Yoon, S., Raghavan, N., Hwang, D., Basnet, S., & Park, H. (2021). Function score-based technological trend analysis. *Technovation*, 101, 102199.
- Nayak, A. S., Kanive, A. P., Chandavekar, N., & Balasubramani, R. (2016). Survey on pre-processing techniques for text mining. *International Journal of Engineering and Computer Science*, 5(6), 16875-16879.
- Nazarenko, A., Vishnevskiy, K., Meissner, D., & Daim, T. (2022). Applying digital technologies in technology roadmapping to overcome individual biased assessments. *Technovation*, 110, 102364.
- Nie, B., & Sun, S. (2017). Using text mining techniques to identify research trends: A case study of design research. *Applied Sciences*, 7(4), 401.
- Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 25). San Francisco, CA,

USA: Determination press.

NLP Stanford. (2022, May 05) GloVe: Global Vectors for Word Representation  
<https://nlp.stanford.edu/projects/glove/>

Nuortimo, K., & Härkönen, J. (2018). Opinion mining approach to study media-image of energy production. Implications to public acceptance and market deployment. *Renewable and Sustainable Energy Reviews*, 96, 210-217.

Ozansoy Çadircı, T., & Sağkaya Güngör, A. (2021). 26 years left behind: a historical and predictive analysis of electronic business research. *Electronic Commerce Research*, 21(1), 223-243.

Ozcan, S., Homayounfard, A., Simms, C., & Wasim, J. (2021). Technology roadmapping using text mining: A foresight study for the retail industry. *IEEE Transactions on Engineering Management*, 69(1), 228-244.

Pandelea, V., Ragusa, E., Apicella, T., Gastaldo, P., & Cambria, E. (2021). Emotion Recognition on Edge Devices: Training and Deployment. *Sensors*, 21(13), 4496.

Park, H., Bharadhwaj, H., & Lim, B. Y. (2019, July). Hierarchical multi-task learning for healthy drink classification. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Park, G., Nan, D., Park, E., Kim, K. J., Han, J., & del Pobil, A. P. (2021, January). Computers as social actors? Examining how users perceive and interact with virtual influencers on social media. In *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)* (pp. 1-6). IEEE.

Park, H., Phaal, R., Ho, J. Y., & O'Sullivan, E. (2020). Twenty years of technology and strategic roadmapping research: A school of thought perspective. *Technological Forecasting and Social Change*, 154, 119965.

- Parwez, M. A., & Abulaish, M. (2019). Multi-label classification of microblogging texts using convolution neural network. *IEEE Access*, 7, 68678-68691.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). " Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, p.
- Pejic-Bach, M., Bertoncel, T., Meško, M., & Krstić, Ž. (2020). Text mining of industry 4.0 job advertisements. *International journal of information management*, 50, 416-431.
- Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
- Pitt, C., Park, A., & McCarthy, I. P. (2021). A bibliographic analysis of 20 years of research on innovation and new product development in technology and innovation management (TIM) journals. *Journal of Engineering and Technology Management*, 61, 101632.
- Pollack, C. C., Gilbert-Diamond, D., Alford-Teaster, J. A., & Onega, T. (2021). Language and Sentiment Regarding Telemedicine and COVID-19 on Twitter: Longitudinal Infodemiology Study. *Journal of medical Internet research*, 23(6), e28648.
- Porter, A., Youtie, J., Carley, S., Newman, N., & Murdick, D. (2018, September). Contest: Measuring tech emergence. In *STI 2018 Conference Proceedings* (pp. 1440-1442). Centre for Science and Technology Studies (CWTS).
- Qian, Y., Ni, Z., Gui, W., & Liu, Y. (2021). Exploring the Landscape, Hot Topics, and Trends of Electronic Health Records Literature with Topics Detection and Evolution Analysis. *Int. J. Comput. Intell. Syst.*, 14(1), 744-757.
- Qiu, Z., & Wang, Z. (2020). Technology forecasting based on semantic and citation analysis of patents: A case of robotics domain. *IEEE Transactions on Engineering*

## *Management.*

- Rahman, S. R., Islam, M. A., Akash, P. P., Parvin, M., Moon, N. N., & Nur, F. N. (2021). Effects of co-curricular activities on student's academic performance by machine learning. *Current Research in Behavioral Sciences*, 2, 100057.
- Rashid, R. A., Chin, L., Sarijari, M. A., Sudirman, R., & Ide, T. (2019, July). Machine learning for smart energy monitoring of home appliances using IoT. In *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)* (pp. 66-71). IEEE.
- Robinson, D. K., Huang, L., Guo, Y., & Porter, A. L. (2013). Forecasting Innovation Pathways (FIP) for new and emerging science and technologies. *Technological Forecasting and Social Change*, 80(2), 267-285.
- Rong, X. (2014). word2vec parameter learning explained. arXiv preprint arXiv:1411.2738.
- Roy, D., Ganguly, D., Bhatia, S., Bedathur, S., & Mitra, M. (2018, October). Using word embeddings for information retrieval: How collection and term normalization choices affect performance. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 1835-1838).
- Roy, K. C., Hasan, S., & Mozumder, P. (2020). A multilabel classification approach to identify hurricane-induced infrastructure disruptions using social media data. *Computer-Aided Civil and Infrastructure Engineering*, 35(12), 1387-1402.
- Rane, A., & Kumar, A. (2018, July). Sentiment classification system of twitter data for US airline service analysis. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 1, pp. 769-773). IEEE.
- Rastegari, E., & Ali, H. (2020). A bag-of-words feature engineering approach for assessing health conditions using accelerometer data. *Smart Health*, 16, 100116.

- Rathan, M., Hulipalled, V. R., Venugopal, K. R., & Patnaik, L. M. (2018). Consumer insight mining: aspect based Twitter opinion mining of mobile phone reviews. *Applied Soft Computing*, 68, 765-773.
- Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*.
- Ren, J., Long, J., & Xu, Z. (2019). Financial news recommendation based on graph embeddings. *Decision Support Systems*, 125, 113115.
- Ren, H., & Zhao, Y. (2021). Technology opportunity discovery based on constructing, evaluating, and searching knowledge networks. *Technovation*, 101, 102196.
- Rezaeinia, S. M., Rahmani, R., Ghodsi, A., & Veisi, H. (2019). Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications*, 117, 139-147.
- Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, 8(627-633), 116.
- Rosenberg, A., & Hirschberg, J. (2007, June). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (pp. 410-420).
- Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology?. *Research policy*, 44(10), 1827-1843.
- Sakahira, F., & Hiroi, U. (2021). Designing cascading disaster networks by means of natural language processing. *International Journal of Disaster Risk Reduction*, 66, 102623.



- Saura, J. R. (2021). Using data sciences in digital marketing: Framework, methods, and performance metrics. *Journal of Innovation & Knowledge*, 6(2), 92-102.
- Saura, J. R., & Bennett, D. R. (2019). A three-stage method for data text mining: Using UGC in business intelligence analysis. *Symmetry*, 11(4), 519.
- Schemmann, B., Herrmann, A. M., Chappin, M. M., & Heimeriks, G. J. (2016). Crowdsourcing ideas: Involving ordinary users in the ideation phase of new product development. *Research Policy*, 45(6), 1145-1154.
- Shahid, N., Rappon, T., & Berta, W. (2019). Applications of artificial neural networks in health care organizational decision-making: A scoping review. *PloS one*, 14(2), e0212356.
- Shahid, F., Zameer, A., & Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons & Fractals*, 140, 110212.
- Shenfield, A., Day, D., & Ayes, A. (2018). Intelligent intrusion detection systems using artificial neural networks. *Ict Express*, 4(2), 95-99.
- Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).
- Silva, F. B., Werneck, R. D. O., Goldenstein, S., Tabbone, S., & Torres, R. D. S. (2018). Graph-based bag-of-words for classification. *Pattern Recognition*, 74, 266-285.
- Simon, H. (2009). Neural networks and learning machines.
- Skansi, S. (2018). *Introduction to Deep Learning: from logical calculus to artificial intelligence*. Springer.

- Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research policy*, 43(8), 1450-1467.
- Smilkov, D., Thorat, N., Nicholson, C., Reif, E., Viégas, F. B., & Wattenberg, M. (2016). Embedding projector: Interactive visualization and interpretation of embeddings. arXiv preprint arXiv:1611.05469.
- Song, G., & Huang, D. (2021). A sentiment-aware contextual model for real-time disaster prediction using Twitter data. *Future Internet*, 13(7), 163.
- Song, K., Kim, K. S., & Lee, S. (2017). Discovering new technology opportunities based on patents: Text-mining and F-term analysis. *Technovation*, 60, 1-14.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929-1958.
- Sun, X., Han, M., & Feng, J. (2019). Helpfulness of online reviews: Examining review informativeness and classification thresholds by search products and experience products. *Decision Support Systems*, 124, 113099.
- Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D. (2020). Mobilebert: a compact task-agnostic bert for resource-limited devices. arXiv preprint arXiv:2004.02984.
- Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text mining: techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, 7(11), 414-418.
- Tang, K. Y., Chang, C. Y., & Hwang, G. J. (2021). Trends in artificial intelligence-supported e-learning: A systematic review and co-citation network analysis (1998–2019). *Interactive Learning Environments*, 1-19.
- Team, T. (2020). Pandas development Pandas-dev/pandas: Pandas. *Zenodo*, 21, 1-9.

- Tellez, A., Pumperla, M., & Malohlava, M. (2017). *Mastering Machine Learning with Spark* 2. x. Packt Publishing Ltd.
- Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. arXiv preprint arXiv:1905.05950.
- Thakur, K., & Kumar, V. (2020). An Overview of Text Mining: Application and Free Software Tools. *Library Waves*, 6(2), 53-59.
- Thangaraj, M., & Amutha, S. (2018). Mgephi: Modified gephi for effective social network analysis. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, 1(1), 39-50.
- Tiwari, M., Bharuka, R., Shah, P., & Lokare, R. (2020). Breast cancer prediction using deep learning and machine learning techniques. Available at SSRN 3558786.
- Trott, P. (2001). The role of market research in the development of discontinuous new products. *European Journal of Innovation Management*.
- UHeM. (2022, May 24) Technical Specifications of Server <https://en.uhem.itu.edu.tr/donanim.html>
- Vahidnia, S., Abbasi, A., & Abbass, H. A. (2021). Embedding-based Detection and Extraction of Research Topics from Academic Documents Using Deep Clustering. *Journal of Data and Information Science*, 6(3), 99-122.
- Van Eck, N. J., & Waltman, L. (2013). VOSviewer manual. Leiden: Univeristeit Leiden, 1(1), 1-53.
- Van Rossum, G. (2020). *The Python Library Reference, release 3.8.2*. Python Software Foundation.
- Verma, S., & Gustafsson, A. (2020). Investigating the emerging COVID-19 research trends in the field of business and management: A bibliometric analysis approach. *Journal of*

*Business Research*, 118, 253-261.

- Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11, 2837-2854.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & Van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, 17(3), 261-272.
- von Delft, S., & Zhao, Y. (2021). Business models in process industries: Emerging trends and future research. *Technovation*, 105, 102195.
- Wajahat, A., Nazir, A., Akhtar, F., Qureshi, S., Razaque, F., & Shakeel, A. (2020, January). Interactively visualize and analyze social network Gephi. In *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (pp. 1-9). IEEE.
- Wang, Y. Y., Chi, Y. Y., Xu, J. H., & Li, J. L. (2021). Consumer Preferences for Electric Vehicle Charging Infrastructure Based on the Text Mining Method. *Energies*, 14(15), 4598.
- Wang, J. H., Liu, T. W., Luo, X., & Wang, L. (2018, October). An LSTM approach to short text sentiment classification with word embeddings. In *Proceedings of the 30th conference on computational linguistics and speech processing (ROCLING 2018)* (pp. 214-223).
- Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.
- Wei, N. J., Dougherty, B., Myers, A., & Badawy, S. M. (2018). Using Google Glass in surgical settings: systematic review. *JMIR mHealth and uHealth*, 6(3), e9409.

- Wei, W., Nan, D., Zhang, L., Zhou, J., Wang, L., & Tang, X. (2020, October). Short text data model of secondary equipment faults in power grids based on LDA topic model and convolutional neural network. In *2020 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC)* (pp. 156-160). IEEE.
- Weiss, S. M., Indurkha, N., Zhang, T., & Damerau, F. (2010). *Text mining: predictive methods for analyzing unstructured information*. Springer Science & Business Media.
- Wikipedia. (2022, January 29) Biological neuron model  
[https://en.wikipedia.org/wiki/Biological\\_neuron\\_model](https://en.wikipedia.org/wiki/Biological_neuron_model)
- Wilk, V., Cripps, H., Capatina, A., Micu, A., & Micu, A. E. (2021). The state of digital entrepreneurship: a big data Leximancer analysis of social media activity. *International Entrepreneurship and Management Journal*, 17(4), 1899-1916.
- Xiao, D., Wang, Q., Cai, M., Zhu, Z., & Zhao, W. (2019). A3ID: an automatic and interpretable implicit interference detection method for smart home via knowledge graph. *IEEE Internet of Things Journal*, 7(3), 2197-2211.
- Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation*, 31(7), 1235-1270.
- Yun, S., Song, K., Kim, C., & Lee, S. (2021). From stones to jewellery: Investigating technology opportunities from expired patents. *Technovation*, 103, 102235.
- Zeng, M. A. (2018). Foresight by online communities—The case of renewable energies. *Technological Forecasting and Social Change*, 129, 27-42.
- Zhang, Y., Chen, H., Lu, J., & Zhang, G. (2017). Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016. *Knowledge-Based Systems*, 133, 255-268.
- Zhou, R., Awasthi, A., & Stal-Le Cardinal, J. (2021). The main trends for multi-tier supply

chain in Industry 4.0 based on Natural Language Processing. *Computers in Industry*, 125, 103369.

Zhou, Y., Dong, F., Liu, Y., Li, Z., Du, J., & Zhang, L. (2020). Forecasting emerging technologies using data augmentation and deep learning. *Scientometrics*, 123(1), 1-29.

Zou, W., Sastry, M., Gooding, J. J., Ramanathan, R., & Bansal, V. (2020). Recent advances and a roadmap to wearable UV sensor technologies. *Advanced Materials Technologies*, 5(4), 1901036.

