Emerging Trend Detection in News Articles

Nhu Khoa Nguyen

LA ROCHELLE UNIVERSITÉ

École doctorale Euclide

Laboratoire Informatique, Image, Interaction (L3i)

THÈSE présentée par :

Nhu Khoa NGUYEN

soutenance le : **27 février 2023**

pour obtenir le grade de : **Docteur de La Rochelle Université**

Discipline : **Informatique et Applications**

Emerging Trend Detection in News Articles

Rapporteurs	**Anne VILNAT**	Professeure des Universités, LISN, Université Paris-Saclay
	Fabrice MAUREL	Maître de Conférences HDR, GREYC, Université de Caen Normandie
Examinateurs	**Gaël DIAS**	Professeur des Universités, GREYC, Université Caen Normandie
	Karell BERTET	Professeure des Universités, L3i, La Rochelle Université
Directeurs	**Antoine DOUCET**	Professeur des Universités, L3i, La Rochelle Université
	Thierry DELAHAUT	La Banque Postale Asset Management
Co-encadrants	**Gaël LEJEUNE**	Maître de Conférences, STIH, Sorbonne Université
	Emanuela BOROS	Collaboratrice Scientifique, Ecole Polytechnique Fédérale de Lausanne (EPFL)


La Rochelle
Université

Abstract

In the financial domain, information plays an utmost important role in making investment/business decisions as good knowledge can lead to crafting correct approaches in how to invest or if the investment is worth it. Moreover, being able to identify potential emerging themes/topics is an integral part of this field, since it can help get a head start over other investors, thus gaining a huge competitive advantage. To deduce topics that can be emerging in the future, data such as annual financial reports, stock market, and management meeting summaries are usually considered for review by professional financial experts. Reliable sources of information coming from reputable news publishers, can also be utilized for the purpose of detecting emerging themes. Unlike social media, articles from these publishers have high credibility and quality, thus when analyzed in large sums, it is likely to discover dormant/hidden information about trends or what can become future trends. However, due to the vast amount of information generated each day, it has become more demanding and difficult to analyze the data manually for the purpose of trend identification.

Our research explores and analyzes data from different quality sources, such as scientific publication abstracts and a provided news article dataset from Bloomberg called Event-Driven Feed (EDF) to experiment on Emerging Trend Detection. Due to the enormous amount of available data spread over extended time periods, it encourages the use of contrastive approaches to measuring the divergence between past and present surrounding context of extracted words and phrases, thus comparing the similarity between unique vector representations of each interval to discover movement in word usage that can lead to the discovery of new trend. Experimental results reveal that the assessment of context change through time of selected terms is able to detect critical emerging trends and points of emergence. It is also discovered that assessing the evolution of context over a long time span is better than just contrasting the two most recent points in time.

Résumé

Dans le domaine de la finance, l'information joue un rôle extrêmement important dans la prise de décisions en matière d'investissement. En effet, une meilleure connaissance du contexte peut conduire à l'élaboration d'approches plus appropriées quant à la manière d'investir et à la valeur de l'investissement. En outre, être capable d'identifier les thématiques émergentes fait partie intégrante de ce domaine, car ceci peut aider à prendre de l'avance sur les autres investisseurs, et donc à obtenir des avantages concurrentiels considérables. Pour identifier les thèmes susceptibles d'émerger à l'avenir, des sources telles que les rapports financiers annuels, les données des marchés boursiers ou encore les résumés des réunions de la direction sont examinées par des experts financiers professionnels. Des sources d'information fiables provenant d'éditeurs de presse réputés peuvent également être utilisées pour détecter les thèmes émergents. Contrairement aux médias sociaux, les articles de ces éditeurs jouissent d'une crédibilité et d'une qualité élevées. Ainsi, lorsqu'ils sont analysés en grande quantité, il est probable que l'on découvre des informations dormantes/cachées sur les tendances ou ce qui peut devenir des tendances futures. Cependant, en raison de la grande quantité d'informations générées chaque jour, il est devenu plus exigeant et difficile d'analyser les données manuellement tout en détectant les tendances au plus vite.

Notre recherche explore des données de différentes sources de qualité, telles que des résumés de publications scientifiques et un ensemble de données d'articles d'actualité fournis par Bloomberg, appelé Event-Driven Feed (EDF), afin d'expérimenter la détection des tendances émergentes. En raison de l'énorme quantité de données disponibles réparties sur de longues périodes de temps, elle encourage l'utilisation d'une approche contrastive pour mesurer la divergence entre le contexte environnant, passé et présent des mots et des phrases extraits, comparant ainsi la similarité entre les représentations vectorielles uniques de chaque intervalle pour découvrir des évolutions dans l'utilisation des termes qui peuvent conduire à la découverte d'une nouvelle tendance émergente. Les résultats expérimentaux révèlent que l'évaluation de l'évolution dans le temps du contexte des termes est susceptible de détecter les tendances critiques et les points d'émergence. On découvre également que l'évaluation de l'évolution du contexte sur une longue période est préférable à la simple comparaison des deux points les plus proches dans le temps.

Acknowledgement

The thesis was the biggest challenge in my life thus far, with many ups and downs, moving to a new country, and exploring a new field of research while in the midst of a global epidemic. First and foremost, I would like to express my deepest gratitude to my two supervisors, Professor Antoine DOUCET and Doctor Thierry DELAHAUT, for their immense knowledge and continuous support, both on the academic side and the administrative side. Thank you so much for having patience with me at the beginning of the thesis when things did not go as planned due to the COVID-19 restriction.

Besides my two supervisors, I would also like to thank Assoc. Prof. Gaël LEJEUNE and Doctor Emanuela BOROS for co-supervising this thesis. Together with Antoine and Thierry, they constantly challenged and pushed me and my work toward the best possible version I can achieve while understanding and supporting me during times when I was having a slump and had little motivation. I could not express with words how grateful I am for everyone to have faith in me to push past those difficult phases.

I am thankful also to my doctoral committee for accepting to put valuable time in reading, reviewing, and challenging ideas presented in my thesis with unique points of view, Gaël DIAS, Anne VILNAT, Fabrice MAUREL, and Karell BERTET.

The thesis was also a unique opportunity for me to carry out my research in an industrial setting. For this, I would like to thank La Banque Postale - Asset Management in general and the Smart Beta team specifically for this once-in-a-lifetime chance. Being able to conduct research through the lens of the finance industry has helped me get a better and broader outlook on both academic and professional points of view. Thank you to my colleagues on the Smart Beta team, Robin, Caroline, Olivier, and especially Stephane for helping me during my time at the La Banque Postale - Asset Management.

Even though my time at the L3i Laboratory was short and infrequent, mostly due to COVID-19, I would like to thank my lab mates for their warm welcome every time I returned, for sharing conversations and experiences to remind myself that I am not alone in this journey, for helping me in times of need: Tien Nam, Tri Cong, Beatrix, Viviana and many more.

I would also want to express my gratitude to the University of Science and Technology of Hanoi (USTH) as well as the ICT department and their lectures for laying the foundation that helped me become who I am today and achieved this milestone. Thanks to all the friends I have made at USTH that listened to my troubles, helped me become a better person, and kept me level-headed in times of hardship: Hoang Minh, Lam Dang, Duc Thang, Nhat Linh and Do Hoang to name a few.

I would like to thank my parents for their unconditioned love and support during these three years of doing the thesis far away from home.

Finally, I would like to warmly thank my girlfriend, Lan Anh, for her unwavering support throughout the thesis, especially during those difficult periods.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

The amount of information available online grows greatly each passing day, with the ease of accessibility to social media, news portals, journals, etc. Thus, as an effect, it is gradually more demanding and requires domain knowledge and experts to update on the latest events and developments as well as follow trends fading in and out of the public's attention (Goorha and Ungar, 2010). Trends can have a long lifespan or last only a short time frame depending on the nature of the text that promotes it. For example, outburst topics on social media can appear suddenly with a great number of occurrences due to their spontaneous nature of being an information stream (Naaman et al., 2011), while topics reported in newspapers or new research approaches can take a longer time to develop (Cataldi et al., 2014). The excess volume of data that people could not and probably would not process, given limited time in a day, more than often contains latent information, signifying an emergence of a new topic that has the potential to evolve into a trend in the near future. Subsequently, while opinions from experts regarding

trend prediction are still valuable, the ability to automatically process an enormous volume of data and promptly recognize and recommend new emerging topics has become a tremendous advantage. Thus, the situation demands the task of Emerging Trend Detection (ETD), which often intertwines with the Topic Detection and Tracking (TDT) task (Fiscus et al., 1999; Fiscus and Doddington, 2002).

Trends that emerge from news articles can have a high impact on many aspects. For example, when the H1N1 disease broke out in 2009 in Germany, news publishers played an important role in raising awareness of how dangerous the virus was by having risk communication as the main focus. The result was an increase in health-promoting behaviour which help to prevent the spread of the disease. Moreover, Husemann and Fischer (2015) found that the rise in the number of H1N1-related news is proportionate with the decrease in infected cases in Germany. On social media, gathering trends can help discover the opinion of the public on current governance events, such as the emergence of the healthcare reform debate (with the term "HCR") on Twitter[1], a popular social media platform, on May 25, 2010, which can help with the analysis and policy adjustment from the government (Naaman et al., 2011). In a more recent example, the COVID-19 epidemic has caused a global health crisis that required the government to impose quarantine and lockdown to prevent the spread of infection. During this period, social media has been at the forefront for the public to express their concerns over the disease and opinions on the policy enforced by the government, which established trends regarding public health and quality of life within the community (Chang et al., 2021; Chen et al., 2020; De Santis et al., 2020; Li et al., 2020b). In the case of scientific publication, the prime example is cloud computing in information technology which started in the 1960s and has much-related research since (Kaur and Kaur, 2014). The impact of this research topic has been wide-reaching as many other domains, namely education, health, and social media, have benefited from its applications (Senyo et al., 2018), largely due to major service providers such as Amazon, Oracle, Microsoft, etc. who saw the potential and adopted the technology (Navas Khan, 2020).

The definition of ETD varies based on the scope, objective, and provided data. In its earlier development, a first definition was given, which classified the task as the identification of topic

[1] https://twitter.com

areas of growing interest or utility over time (Kontostathis et al., 2004). As research progressed, the task became more and more diverse due to the variety of data available (Boyack and Klavans, 2022; Cataldi et al., 2010; Moiseeva and Schütze, 2020).

For example, in the context of social media, Cataldi et al. (2010) considered an emerging topic as a set of terms that semantically correlate and are similar to an emerging keyword. The authors described a term as emergent when it was used extensively at a given time period, yet not in previous ones (Cataldi et al., 2010). In another study conducted by Naaman et al. (2011) about Twitter, the research described an emerging topic being just one or more keywords within a time interval that had activities (e.g., the message volume containing said term) exceed expectations with respect to past periods or other terms (Naaman et al., 2011). This research also gave insight into an informal definition given by Twitter, one of the most popular social media platforms, which is "keywords that happen to be popping up in a bunch of tweets". In more recent research regarding trend detection in micro-blogs, Dang et al. (2016) further explained the concept of an "emerging topic", describing it as content that can attract attention in a limited time period, yet could have a long-lasting effect with significantly more influential discussions than common topics and often involve with impactful emerging events such as natural disasters, election campaigns, regulations enforcement, etc.

Early detection of emerging topics yields major advantages in many areas, whether the would-be-emerged topic is short-lived or has a long lifespan. Within the scientific community, the definition of emerging topics is slightly different in that the number of publications surrounding such topics increases significantly within a certain time interval (Moiseeva and Schütze, 2020). Another perspective is that a publication in a high-impact journal can trigger an increase in attention toward it and the related research field (He and Chen, 2018). Thus, identifying potential research trends in advance is useful for preparing to fund and allocate resources toward promising projects (Behpour et al., 2021a). Trends discovery in research publications can also have a great effect on the development of technology, economy, and society since it can bring more attention from industries, policymakers, and governments to studies with potentially high impact on profit and innovation (Boyack and Klavans, 2022). For competitive intelligence, a domain in corporate business that thrives on favourable information to gain edges over competitors, iden-

tifying trendsetters as soon as possible is a desirable objective to anticipate future developments and changes that can lead to new and better business strategies (Akrouchi et al., 2021). For example, in the medical field, one of the notable applications of ETD is finding emergencies and disappearances of therapy methods through clinical data analysis to reflect the quality of treatment for further improvement, enhancing the quality of life (Chen et al., 2007b).

In the context of this thesis, in the field of finance, the impact of EDT is immense since being able to detect short-term trending topics can help with the daily stock market decisions and strategy, as new information arises and is reflected through the movement of stock prices rapidly (Prachyachuwong and Vateekul, 2021). Additionally, predicting long-lasting, sustainable emerging themes will aid the process of investing, in which one good investment opportunity can turn into profits in the future.

These aforementioned domains (e.g., scientific, financial, medical, social media) differ from each other in terms of text characteristics, such as density, vocabulary, language formality, and type of information they convey which is either fact, theory, or opinion/emotion/feeling. For example, scientific publications are written in formal language, with academic vocabulary that introduces facts of previous works as well as theories and methods for current research. Documents of financial and medical domains, while having the same formal writing, pay more attention to facts and report current events or findings. In contrast, social media tends to have informal language, including slang and abbreviation in attempts to express one's opinions and beliefs in a more restricted format of a few hundred characters or less than a hundred words.

In recent years of ETD research, social media played a crucial part in the development of the task by having data readily available and enormous in size. This was possible due to the fact that such type of data is able to benefit from metadata that contains information about the data itself. Some elements of the metadata are of great help to the ETD task, such as retweets, likes, and hashtag count for Twitter data, citations, and references number for scientific publications. These metrics indicate the diffusion rate of information, which can directly translate into how popular a topic has become. Nevertheless, many studies, while capitalizing on such diffusion features, did not concentrate on analysing the content of the text as one of the major components of the task, which is most viable for short-text data, yet can cause loss of context in the longer format.

Moreover, while it has become more common for data to have metadata in them, diffusion features as mentioned above are not always available, which can be a liability. Meanwhile, in the context of EDT in financial or medical-related documents, this type of metadata or diffusion feature is lacking, thus increasing the difficulty of detecting emerging trends.

1.2 Objectives for Detecting Trends in Heterogeneous Information

This thesis under a CIFRE contract[2] is a collaboration between “La Banque Postale Asset Management (LBP-AM)” and the University of La Rochelle (L3i laboratory). LBP-AM is a finance-focused company whose expertise is in managing assets of multiple types such as bonds, insurance, debt fund, European equities, etc. Within the core staff of LBP-AM, the quantitative management division “Smart Beta” aims to have strong development in methodology regarding big data and artificial intelligence. One of the goals the team settles on is to identify emerging themes worthy of investment, with emphasis on timeliness, in order to gain a competitive advantage over other investors. Having the same direction with the aforementioned motivations, the LBP-AM opted for experimenting with the Event Driven Feeds (EDF) from Bloomberg, a collection of every news article that the publisher has uploaded from 2009 up until the present. With a wide coverage of many different aspects of society, finance, economy, technology impact, etc., the “Smart Beta" team wished to explore and capture emerging trends potentially hidden in this enormous collection of information. Thus, this raises a need for a robust system that can handle textual data from many fields.

Based on the research motivations both from the scientific aspect and industrial aspect per specifications of the LBP-AM, we define the task of emerging trend detection as the identification of a single keyword or a pair of keywords that have their surrounding context change significantly enough due to a new surge of attention, thus causing a new topic of discussion to emerge. Ultimately, the detected keywords should be closely related to movements that are on the horizon of being the latest development or are gaining more attention than in the past in technology, econ-

[2]CIFRE contracts (Industrial Agreements for Training through Research) are agreements under which a company recruits a PhD student doing doctoral research in collaboration with a public laboratory.

omy, society, etc. Given the impact of said aspects in business and development, these themes and topics should be able and are expected to last a few years in order to be sustainable enough for long-term investment.

Hence, the main objective of this thesis is to detect meaningful emerging themes from high-quality sources of information, e.g. reputable news publishers, with a robust approach that is able to process data from a variety of fields and sources. The dataset that the company mainly wishes to explore is the Event Driven Feeds (EDF) from Bloomberg. The detected emerging themes will assist in finding companies that are for good investment opportunities. Companies that innovate or are early followers of emerging themes were known to do well financial-wise, and thus would be profitable to invest in.

Since analysing and assessing potential investment themes is currently done manually by the "Smart Beta" team by combining custom-built knowledge graphs using the knowledge and experience of team members along with document filtering and keyword searching for creating investment portfolios, this research aims to produce a system that can aid them in the process. Transitioning from a purely human-based operation, the system will return a list of suggested promising topics, which are then evaluated by experts from the team. We decided to follow this semi-automatic system approach because we strongly believe that while the system can process a large volume of data to distil the information, knowledge, and experience from specialists are still necessary for the final forecast. Hence, a good balance between automation and humans is needed where the automated system should provide enough information and context so that it can complement prior valuable knowledge and experience from experts in making investment decisions.

With these goals in mind, we approach the emerging theme detection task as a topic detection task with a time element for tracking changes in the surrounding context of detected topics within the different time windows, with the heterogeneity of data taken into consideration. As established earlier, trends from scientific research can also become investment opportunities. Hence, not only are we focusing on the provided EDF data, but also on scientific publications. The diverse and domain-specific datasets have led the research to have a variety of similar tasks that aim at exploring the text for keywords, and more importantly, assessing their potential

emergence as trends. To achieve this, we consider tracking the dynamics of similarities between pairs of keywords, their past representations, and thus their respective evolution over time. Earlier research has discovered that terms that are used frequently through time rarely change semantically and polysemous terms can have drastic semantic change due to having more diverse context in usage (Hamilton et al., 2016). Hence, context change surrounding a term can potentially signify its emergence in the public's attention due to having their context shifted compared to the previous time span. Finally, in order to validate our approach to the problem, we entered an evaluation campaign called FinSim[3], which is an annual event with shared tasks aimed at practical financial problems.

1.3 Challenges in Trend Detection from Heterogeneous Data

The stated objectives of the research come with the following challenges that need to be addressed. First and foremost, the heterogeneous nature of our selected data leads to the ETD task being domain-specific. Thus, the approach to this task can vary differently depending on the characteristic of the data. Consequently, the difficulty level of the task can potentially increase as the formulate of a robust, domain-adaptable or even domain-dependant approach would be needed to not only suit the scope of our research but also accommodate the vast amount of data and field of interest for investment the LBP-AM has.

The second challenge in our research is the lack of annotated data for our specific task, as well as the insufficient ETD-related metadata at our disposal. At the time of writing, to the best of our knowledge, no annotated dataset that targets directly at the task of ETD is publicly available. Moreover, obtained datasets only provide metadata describing abstract topics related to the text without any clear indication of the specific time of emergence, what has emerged or what has fallen out of popularity. Because of these inadequacies, the task cannot be considered a supervised, easy-to-evaluate problem, thus requiring either an unsupervised approach or adapting alternative methods.

Having diverse, domain-specific datasets invokes another problem in terms of vocabulary, as

[3] https://sites.google.com/nlg.csie.ntu.edu.tw/finweb2021/shared-task-finsim-2

each dataset can contain a domain-specific glossary. For example, financial documents have specific financial terms such as “bond", “capital", or “liquidation". Social media, in contrast, has an abundant amount of special characters, non-standard terms and/or informal abbreviations. These domain-specific vocabularies, while being one of the unique characteristics of a dataset, contribute little to the discovery of topics and trends as they appear quite often and can be treated as noise in the text. Thus, analyzing the impact of noise and pre-processing in each dataset is desired to provide insight into how domain-specific vocabulary affects the end result of the ETD task.

In ETD, the element of time plays an important role in the research, as not having a tolerable timeliness detection can reduce the effectiveness of a method in practical scenarios, hindering its application. Hence, utilizing and making the most out of the temporality factor is another challenge the research needs to resolve. Moreover, since a robust, text-centric approach is desired to deal with a variety of data, and not rely on metadata that is not always available, the research’s main focus is to solve how to integrate temporal factors when analysing the content of the text in a large corpus, spread in a fixed time interval.

Lastly, as stated in the objectives, keywords of interest are the fundamental textual feature that the research will focus on. However, keywords, by themselves, are not particularly useful as they represent only a concept at a certain granularity. Hence, linking the detected emerging keywords into a grander theme is essential to formulate an interpretable topic. This consitutes another challenge that needs to be tackled in order to satisfy the practical objectives and requirements of the LBP-AM.

1.4 Research Questions

Essentially, to articulate and support the research as well as have a clear vision of our challenges, we raise the following research questions in the upcoming sections:

1. How does the trend definition vary and how diverse is the data?
2. Since data is heterogeneous, to what extent do noise and data pre-processing impact the detection of trends?

3. Does context evolution of keywords provide a more timely signal than text volume metric for highlighting emerging trends in heterogeneous data?

4. Are our proposed methods efficient in a real use case scenario?

1.4.1 Research Question 1: How does the trend definition vary and how diverse is the data?

One unique aspect of our research is that we are not only focused on experimenting and implementing methods for just one dataset but multiple datasets from different types of sources and about various domains. Thus, the content of the text in our diverse data, as well as their characteristics can overall affect how we approach our ETD task.

The first question we are going to explore is how diverse is the data and whether the definition of a trend varies depending on it. Chapter 2 presents a literature review on the state of the art in ETD where we stratify existing approaches into different categories. Moreover, Chapter 3 introduces in detail the datasets that are available to us and how they differ.

1.4.2 Research Question 2: Since data is heterogeneous, to what extent do noise and data pre-processing impact the task of ETD?

A noticeable difference between social media data and news articles is the contrast in text length, where the former usually contains around 280 characters while the latter can be a few paragraphs long. The discrepancy in text dictates the nature of context described in each type of data: short-text data tends to be concise to encapsulate as much information as possible, while an article, being longer, has more space to contain more context.

Nevertheless, longer texts are more susceptible to noise and irrelevant information, such as publisher information embedded in the text, articles covering not only one main subject but also a few supporting topics, or tables that could not be converted well into text format. To what degree noisy text affects the performance of Topic Detection is an important question. We will address it in Chapter 4 where we discuss the impact of noise in Topic Detection and adjust our text pre-processing strategy of the EDF data through our findings in studies on health-related

and social media data as well as the news articles corpus provided by the LBP-AM.

1.4.3 Research Question 3: Does context evolution of keywords provide a more timely signal than text volume metric for highlighting emerging trends in heterogeneous data?

As stated in the research objectives, we want to rely as much as possible on the content of the text to detect keywords that can be considered as the centre of an emerging theme to avoid the need to depend on meta-data. Moreover, we would like to compare how context evolution as a feature can help for detecting potential topics, compared to text volume/frequency, since text frequency can be misleading and depends considerably on the scope of the data. Subsequently, the research has a number of challenges that are required to be tackled, most notably data-related challenges due to the heterogeneous nature of available data.

Not only can the usage of terms change over time in terms of frequency, but their context of appearance can also be altered both inversely correlated (law of conformity) or independently to their usage count if used in more diverse scenarios (law of innovation) (Hamilton et al., 2016). Thus, assessing the number of term appearances alone might be inadequate. Additionally, we are interested in selecting the point of reference for tracking context change as it can be argued that for assessing the evolution of context, comparing two consecutive time intervals might be too limited to draw a compelling conclusion.

The question and related points are discussed throughout Chapter 5 and Chapter 6 of the dissertation as we propose our studies on news article collections and scientific publications.

1.4.4 Research Question 4: Are our proposed methods efficient in a real use case scenario?

One of the challenges we mentioned is the lack of annotated dataset to validate our research. Hence, it is certainly up for discussion whether our method of assessing context can be effective in a practical scenario. Moreover, because our approach is set to aim at keyword-level detection, terms that our system selected need to be connected to grander themes in order to make them

understandable and can be elaborated as to why the detected topics can potentially emerge into trends.

Chapter 7 presents the research surrounding this question as we participated in an evaluation campaign for practical financial problems called FinSim. The task we selected at this event was hypernym detection, which has the goal of associating financial terms of finer granularity to their respective abstract concepts, which falls in the context of our research and aligns with the challenge of linking keywords to formulate a more elaborated theme.

In the conclusions (cf. Chapter 8), we reiterate these research questions and summarise their answers according to the results presented in the dissertation. Moreover, limitations and perspectives regarding this research are discussed.

1.5 Contributions of the Thesis

The main contribution of this research concerns a new approach to emerging theme detection in news articles through analyzing the context evolution of terms and keywords representing the topics. In particular, it explores keywords that can have a certain degree of impact within a period of the text corpus, then proceeds to extract their representation and calculates the similarity between past representations of themselves and between different detected keywords to finalize a list of potential topics. This allows experts at the "Smart Beta" team to quickly filter and assess timely information at hand for constructing investment strategies.

The contributions of the dissertation are listed as follows:

1. We formulate the state of the art on ETD by categorizing the approach based on different types of approaches and types of data used. Moreover, due to the nature of our task, as well as the lack of annotated data and evaluation methods, we introduce various datasets that we regard as relevant to our research. These datasets are: a multilingual health reports corpus called DAnIEL, a financial news corpus from a collection called Event-Driven Feeds that Bloomberg provides to LBP-AM, and a computer science publication corpus called TRENDNert. In addition, Twitter-Trend, a collection of text on social media with their respective trends typology, will be briefly mentioned and experimented on.

2. As mentioned previously, the domain-specific datasets available to us have different text characteristics. Because of this, we provide an evaluation of the impact of noise on the Topic Detection task in the case of health reports and extend the study to measure the effect pre-processing can have on the other datasets. Thus, we devise a pre-processing procedure that is best suited for the EDF data from Bloomberg.

3. Using context evolution of selected keywords, we assess emerging trends of different datasets with different approaches to the task. First, we provide a case study of detecting emerging topics of a specific company given the special period of the COVID-19 outbreak using news articles extracted from the Event Driven Feeds. In addition, we introduce the Term Context Evolution series that utilizes context change in different time intervals to track the evolution of context surrounding specific terms and identify terms that have the potential to be trending. This approach is experimented with and applied to the scientific publication corpus TRENDNert.

4. Since the available datasets do not have any annotation to properly assess our approach, we participate in an evaluation campaign and present a method that uses a transformer-based language model with siamese architecture for the task of hypernym extraction with the purpose of connecting keywords to form topics of interest as well as to verify the practicality of our proposed method.

1.6 Thesis Organization

The thesis is organized as follows:

Chapter 2 describes the State of the Art in Emerging Themes Detection regarding different types of methods and different types of data. We primarily focus on approaches that study the actual text to derive potential topics. The five main types of methods that we will cover are the statistical approach, graph-based approach, clustering approach, topic modelling approach, and embedding approach.

In Chapter 3, we present the data that we used in our research, including the data that we adapted our study to in order to obtain a good level of understanding of Emerging Topic Detection and

the Event Driven Feeds from Bloomberg provided by the "Smart Beta" team at LBP-AM.

We address the first research question in Chapter 4 where we study the impact degree of noise on the Topic Detection task. The work considered a system, called DAnIEL, designed for detecting disease outbreaks through journalistic reports that come with an annotated dataset. We concluded errors from this upstream stage are often compounded and propagated to the downstream stages. This leads to our strategy of cleaning and filtering raw data given by the LBP-AM.

Chapter 5 concentrates on the Event Driven Feeds data from Bloomberg, where we apply and adapt methods on a small slice of a grand corpus, focusing on news related to a specific company, Microsoft, in a particular period of the beginning in the COVID-19 outbreak in order to detect emerging topics in this special phase.

In Chapter 6, we introduce a new concept called Term Context Evolution series to track the context change of a given keyword and determine whether the change can be significant enough to consider a keyword to be an emerging topic. This part of the work was conducted on the TRENDNERT dataset (Moiseeva and Schütze, 2020), which is an annotated corpus of Computer Science publications from 1985 to 2015.

We present our work on hypernym detection using siamese Transformers in Chapter 7 which is part of a competition called FinSIM-2 Shared Task 2021 on Learning Semantic Similarities. The task focus on finding broader concepts in terms of meaning with a list of given words in the field of financial (Mansar et al., 2021), which is directly related to the line of work from LBP-AM. The research would help us in connecting popular keywords to form topics/themes in order to decide what can be considered emerging topics.

Finally, Chapter 8 concludes the dissertation with findings and remarks on our work and previously mentioned research questions. In addition, we will also discuss the disadvantages our approach has and propose future works for improvement.

CHAPTER 2

State of the Art

Emerging trend detection from text sources has been attracting significant attention for more than 20 years due to its numerous applications and practicality (Kontostathis et al., 2004). As the literature progresses with different types of textual data being handled, such as news articles (Behpour et al., 2021b), scientific publications (Bolelli et al., 2009a), microblogs or social media (Monselise et al., 2021), etc., approaches to this task became more diverse in order to utilize the specificities of each category, text genre properties such as specific vocabulary, writing structure, or text length for instance, most efficiently (Lucas, 2009).

A text can have less than 300 characters for social media, around 200 words or 1,500 characters for scientific abstract, two-three paragraphs for a news article (accounting for around 600 words), and a few pages that can contain more than eight paragraphs for a journal publication (Hamborg et al., 2018a; Lucas, 2009). Different lengths of text mean the different amounts of information encapsulated could lead to conceiving distinct ETD approaches to yield the best results. In addition to the text length, data can come with supplement features that can track the diffusion of a certain chunk of information, thus having additional factors to identify emerging topics.

In this chapter, we introduce different research works that cover different aspects of ETD. In order to have a better understanding of the approaches presented in the following sections, it is necessary to discuss the two main types of features used in ETD, endogenous and exogenous features (cf. Section 2.1). Endogenous features are, generally, already present within the content of the text, while exogenous features could be generated by using external resources (e.g., language models pre-trained on large amounts of data) or could take the form of metadata, containing statistics on the diffusion rate of the text. Research related to ETD can either use primarily one of the two features or both of them to complement one another.

Afterwards, we discuss two main ways to classify prior works (cf. Section 2.2). First, the literature can be divided into four types of approaches: statistical approaches, graph-based approaches, clustering-based, or topic modelling-based approaches. Second, research methods can be categorized based on the density of text in the data, either sparse text or dense text, for trend detection of certain objectives e.g., long-term trend or short-lived burst of attention (cf. Section 2.3).

2.1 Types of Features

In this section, we present the features commonly used in ETD, which we separated into two main categories: endogenous and exogenous features. These terms are generally utilized in the context of time series forecasting, where the input data can be divided in order to better understand its relationship to the output variable (Vogt and Johnson, 2011). The endogenous input variables are influenced by other variables in the system and on which the output variable depends (e.g., flower growth is affected by sunlight and is therefore endogenous), while the exogenous input variables are those not influenced by other variables in the system and on which the output variable depends (e.g., in a simple causal system such as farming, variables such as weather, farmer skill, pests, and availability of seed are all exogenous to crop production). In the context of NLP, we consider that this categorization is transferable to our task. Thus, endogenous features are those determined by a system (e.g., bag-of-words models or word representations that are trained directly on the corpora per se), and exogenous features are those determined outside the context of the system being imposed on it (i.e., features produced by using external

resources such as language models pre-trained on large amounts of data).

First, we discuss the endogenous features that can be extracted within the content of the text in the context of ETD (e.g., bag-of-words, TF-IDF, LDA topics). Next, we discuss the exogenous features which are based on large pre-trained language models, models that are rather new in research in ETD. Finally, we observe that these two types of features are naturally combined for building efficient ETD systems.

2.1.1 Endogenous Features: Strengths and Limitations

Endogenous features use characteristics within the text itself to form content-based representations. The straightforward method of representing text is to use the bag-of-words model, which segments the text into tokens (i.e., words) and considers the identity of all words within the documents in different n-grams of choice, such as uni-grams and bi-grams (Goldberg and Hirst, 2017). Each unique token can be quantified by counting its occurrences, which will be used to represent the text in the form of a vector of appearances. This feature is often known as term frequency (TF). In many cases, a weighting scheme can be considered to give more importance to less popular words with respect to the corpus compared to more popular ones, which leads to the creation of Inverse Document Frequency (IDF) (Jones, 2004). Combined with TF, the metric became Term Frequency - Inverse Document Frequency (TF-IDF), in which while appearing often is a favourable sign, the importance of a term will be penalized for occurring in too many documents (Robertson, 2004). TF and TF-IDF have been used as a weighting metric to gauge term importance for keywords in the corpus, especially with TF-IDF further improving bag-of-word representation by taking into account terms repetition. Okapi-BM25 (Robertson and Walker, 1994; Robertson et al., 1994), often called Okapi, is another, less popular weighting scheme that is more optimized than TF-IDF, by introducing document length to the equation, thus replacing TF with term saturation.

While bag-of-words features can capture every unit of text in the documents, they have a limitation where all the semantics of the text are lost, in particular since the word order is not retained when extracting features (Naseem et al., 2020). Moreover, the dimension of this kind of representation can become enormous, due to the growing amount of text data at the rate of zettabytes

(ZB) (Swaraj et al., 2015), thus putting heavy stress on computational power. To retain word order, one solution is to use n-gram representation to create a list of words that contain more than one token, which keeps the arrangement of tokens to a certain extent, yet is more computationally intensive. Stemming and lemmatisation are generally known methods used to reduce the representation size where the former preserves the base form of the token in the former by removing common prefixes and suffixes and the latter considers the context in a given lexicon to extract the original word. While the two methods can reduce the number of tokens in the bag-of-words feature, they still do not model semantics well as context can be lost when only the base form of the token is kept.

Word embeddings aimed to solve these problems by representing each token in the text within a pre-defined vector space, where similar tokens should have similar representation in said vector space. The similarity is expressed through their surrounding words as context using distributional hypothesis (Miller and Charles, 1991), with the assumption that the more resemblance the surrounding words are, the closer in terms of meaning the considered words are (Goldberg and Hirst, 2017).

Word2Vec Mikolov et al. (2013a,b), being the most innovative (at the time of publication) and one of the most popular word embeddings algorithms to date, can capture both syntactic and semantic similarities Jiao and Zhang (2021). In later years, GloVE Pennington et al. (2014), which is short for Global Vector for Word Representation, covered the disadvantage of local representation of Word2Vec by incorporating word co-occurrences statistic of the corpus in the process of generating vector, thus creating a global representation of words with respect to the corpus. FastText, created by Bojanowski et al. (2016), is an extension of Word2Vec that focus on the effectiveness of smaller datasets. Instead of feeding n-gram tokens, the algorithm uses n-gram characters for training, thus becoming faster and simpler to train on smaller datasets.

In addition, Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a topic modelling algorithm, can also have its result used as a feature for ETD. Currently, two common approaches exist that utilize LDA output to represent a corpus. Firstly, at the word level, the algorithm output the probability of how suitable an n-gram is at expressing a set number of topics, each n-gram can be represented as a vector of topics probability (Behpour et al., 2021a). The second approach

uses the topic distribution of a document from LDA, where each document is given a probability of how likely it belongs to the set of topics (Xie and Xing, 2013). Thus, documents in the corpus are represented in a vector with the dimension of the number of topics chosen for LDA.

2.1.2 Exogenous Features: Strengths and Limitations

In contrast to endogenous features, exogenous ones rely on external sources of information to redescribe the data. In data mining, re-description is defined as two different ways of describing roughly the same concept, both of which can be of interest (Galbrun and Miettinen, 2017), or one is better described than the other in terms of time/space consistency (Rioult, 2017). The concept also applies to word embeddings mentioned in the previous section as they can be generated by having trained directly on either internal corpus for a designated task or external, more generalized corpora, which produce different representations of the same term. However, while this type of embedding is effective in capturing text syntax and semantics, they are still static vectors, thus having only one global representation for all of their appearances regardless of context (Gupta and Jaggi, 2021). Thus, the lack of consistency in context raised a need for a context-dependent method of generating word embedding, which can be considered as a re-description of the word representation. A Transformer-based language model is composed of self-attention mechanisms and positional embeddings (Vaswani et al., 2017) and is able to learn the context of given training data by generating a language model. While it requires a huge amount of data to train, the result yields a pre-trained language model that is able to output word embeddings dynamically based on the surrounding context. The embeddings generated by pre-trained models using an external corpus can be considered a type of exogenous feature, as text from outside sources gives the pre-trained model the context it needs. Moreover, the model can be further strengthened by fine-tuning with an annotated corpus of choice, thus increasing the quality of text representation.

Also in the exogenous feature category is data on diffusion metrics of text and information. This type of data is generally available in the metadata of the document, quantifying how information is spread. Different types of documents can have their own unique diffusion metrics. For example, in social media, they are a number of reactions and shares of the post (e.g., likes and retweets

accordingly on Twitter specifically). View count is how news sites measure the engagement and exposure of each article (Knobloch-Westerwick et al., 2005). Scientific publication corpus has citation counts which can gauge how impactful an article is to the research community and a references list that can be used to build a graph, connecting related research together (Pestryakova et al., 2022).

2.2 Types of Approaches for Emerging Trend Detection

This section focuses on the various types of approaches toward ETD, using previously mentioned features. Some research relies solely on just one set of features, while others combine two or more sets, or even use both endogenous and exogenous features in their system in order to answer specific questions in ETD regarding the data experimented on. Primarily, we discuss the following categories of methods:

- *Statistical methods* use purely numerical models to extract salient terms/keywords (Hughes et al., 2020).
- *Graph-based methods* establish connections between different levels of text and extract groups of nodes that represent topics (Dang et al., 2016).
- *Clustering approach* centers around grouping text and entities that have similar features to form potential topics of interest (Farzindar and Khreich, 2015).
- *Topic modeling* aims toward using topic probability to discover the text of the same content (Al-Attar and Shaalan, 2021).

2.2.1 Statistical Approaches

Statistical approaches rely on applying statistical techniques to weight terms in order to extract keywords that are important terms within the text to form topics of interest. The most straightforward form of the statistical approach is to find the most frequent terms that exist in the corpus, which considers the Term Frequency feature. However, this approach was shown, through the research of Karen Spärck Jones, as flawed since text that appears frequently contributes more

toward noise than meaningful terms (Jones, 2004). With the introduction of TF-IDF, researchers have found better metrics to represent a term's impact, thus extracting important words within the corpus (Robertson, 2004).

Using TF-IDF as a feature along with WordNet, a large English lexicon database that groups nouns, verbs, adjectives, and adverbs into groups of synonyms(Kohl et al., 2003), Daud et al. (2021) introduces Hot Topic Rising Star (HTRS) system that not only detects hot terms but also explore potential junior authors associating with upcoming research interest regarding emerging keywords. The proposed method utilized TF-IDF to calculate the exact word similarity as the influenced score for ranking hot topics in a scientific publication corpus, which outperformed WordNet semantic similarity ranking in detecting hot topics as well as identifying top contributed authors for said topics. In another research, Fang et al. (2014) applied TF-IDF with extra weight for different types of keywords (for example word, phrase, hashtag) as ranking metrics for keyword extraction and selected top N terms as topic representatives after clustering tweets obtained from Twitter.

While being one of the most popular weighting schemes for extracting terms based on their impact in the corpus, TF-IDF is not without disadvantages. In recent research, Zhu et al. (2019) specified that the metric was not a good text feature to use in hot topic detection due to the fact that in hot topics, new terms appear more frequently in documents, hence lowering the IDF weight. As a consequence, the importance of new terms is lowered. Thus, the authors proposed a refined version of TF-IDF to discover hot terms based on the time distribution of information where an adjusted version of IDF, denoted as $Adjusted_{IDF}$, is introduced, taking into account the IDF value distribution over a period of time. The $Adjusted_{IDF}$ made hot terms more apparent in stages of emergent where such terms are used more frequently and appear in more documents, which mitigated the said drawback of the original IDF. To detect hot topics according to hot terms, K-means clustering algorithm (MacQueen, 1967) is utilized. Experimental results indicated that the refined TF-IDF algorithm can discover hot terms more efficiently, where hot terms were placed ten times higher in terms of importance compared to the original metric, making them more identifiable.

Another drawback of TF-IDF was mentioned in research that studied text from cybersecurity

forum discussion, where Hughes et al. (2020) stated that due to the nature of the data being prone to noise with stopwords and informal language usage, such as slang, acronyms, misspellings, heavy text pre-processing to remove such unwanted text would be needed for TF-IDF to work efficiently. Thus, instead of working with TF-IDF and having to define a list of stopwords and remove them, an approach that focuses on calculating log-odds of terms, which indicate how likely a term to be appeared based on past appearances, was employed, thus capturing linguistic change to deduce trending terms in the corpus. The decision to skip the stopwords removal step was based on a finding in the study, showing that since stopwords usually have similar distribution across any given time period, their log-odds score will be low compared to other text, therefore will not interfere with the actual important information. Experimental results showed that when comparing to TF-IDF, the proposed log-odds approach generates more relevant keywords in both selected corpus about a specific event and corpus with random events.

In another research, Griol-Barres et al. (2020) worked on detecting weak signals in a text corpus of mixed types, including scientific, journalistic, and social media sources on the topic of Remote Sensors. The publication defined weak signal as “a type of future signal that can be defined as evidence of emerging changes into a continuous process of exploration of a specific environment". The signal is hardly recognizable in the current situation, yet might become an important trend that will evolve into a strong signal that can reinforce or hinder existing trends and approaches, which is highly relatable in EDT. Three new metrics are introduced as signals to be evaluated: Degree of Visibility (DoV), Degree of Diffusion (DoD), and Degree of Transmission (DoT). DoT relies on external H-index metrics specifically for scientific publication in the dataset, while DoV and DoD use TF and IDF correspondingly for their calculation. The main strength of the proposed system is that it can be applied to any field with automatic category classification for the detected keywords.

While methods using statistics have shown some success in ETD, aside from the previously mentioned disadvantages, the method still has one glaring limitation: they ignore semantics and context when assessing terms and keywords. Consequently, crucial information could be lost when evaluating raw frequency numbers, hence results can be misleading due to missing context.

2.2.2 Graph-based Approaches

Graph-based approaches in ETD focus on constructing graphs of features in the corpus in order to establish relationships between nodes of events, keywords, documents, etc. They take advantage of such connections to categorize and extract groups of nodes that belong to a certain topic. Connections between nodes can be built from a combination of the following, but not limited to, features: co-occurrences (Ohsawa et al., 1998), temporal features (Glavaš and Šnajder, 2014), and semantic relations (Zhang et al., 2015).

Story Forest is a system that automatically clusters a stream of documents into events and connects relevant events into an unfolding storyline. It was introduced to cover the problem of breaking news coverage (Liu et al., 2020). In their research, the authors constructed a tree to represent a trending topic that is in development through many storylines, with each of them containing multiple events, yet an event cannot belong to multiple topics. In their definition, events are comprised of one or more news articles reporting the same content of real-world breaking news, so long as the content is about the same entities and is published during the same time frame. Having the topics-events-news tree generated, the author also introduced EventX, a 2-layer, a graph-based clustering algorithm that extracts events from enormous news corpora on a finer grain than past methods with multiple levels of representing an event in the constructed tree, clustering topics using keyword co-occurrences graph and extracting events based on document relationship graph. Experimented on the '20 Newsgroups dataset and 60GB of Chinese News corpus, the proposed system was more effective and generated cleaner event clusters when handling unstructured data when compared to existing methods. Also, Story Forest and EventX allow for the dynamic integration of new events into an existing tree, while also being efficient and scalable, which is important for industry practice.

Leskovec et al. (2009) presented a framework that is capable of detecting short-term hot topics news sources by building a graph of distinct phrases, partitioning it into clusters of phrases that are textually similar. Items such as blog posts, news articles and that contain phrases in a cluster of the graph are considered as one thread, and each thread is assessed based on their temporal variation to find their peak of attention. The framework experimental result provides an interesting comparison of peak time and decay rate in attention between blog posts and conventional

news media, where conventional news media can reach the highest point of attention 2.5 hours earlier, yet lose it much faster.

Large-scale approaches that can process enormous amounts of data to identify long-term trends, however, were shown to be inadequate in early detection, which causes the lack of timeliness in the prediction. Dang et al. (2016) explained the main reason was that in the early phase of detection, the text corpus size is small, thus large-scale method could not work effectively due to lack of sample size. Thus, the author proposed an approach based on Dynamic Bayesian Network (DBN) (Murphy, 2002), which is a direct acyclic graph that models conditional dependencies between nodes of variables over adjacent time steps. Each keyword is modelled through 4 observable variables (total number of retweets, number of chain retweets, the maximum size of the retweet chain, and total number of followers) and three hidden variables. Attractiveness is represented by the number of retweets and retweet chains. Key-node relies on a maximum retweet chain and a number of followers. Lastly, emergence depends on the fore-mentioned hidden variables, using a threshold to distinguish between non-emerging and emerging keywords, which are then clustered into topics using DBSCAN (Ester et al., 1996). Experimental results showed that the approach based on DBN can appropriately characterize the temporal evolution of keyword diffusion to identify emerging keywords. Moreover, the proposed system outperformed conventional methods in timeliness, producing results that are averaging about 1.5 hours faster in real-time text stream over the course of 10 hours while maintaining similar detection performance.

To address the problem of utilizing tweets without any hashtags for ETD, Majdabadi et al. (2020) constructed a graph of three main elements of the tweet, word, and hashtag with three types of edge tweet-word, word-hashtag, and tweet hashtag. Edge is calculated using the TF-IDF score and co-occurrences score, where more weight is placed on hashtag relation with a tweet, which ensures that a non-hashtag tweet is considered while a tweet with the hashtag is deemed more influential. By applying clustering and ranking method to the constructed graph, trends can be extracted with a high degree of relevancy and coherency. The performance showed that when compared to classic topic modelling and clustering algorithm, the accuracy of the proposed system is suggested to be superior.

Pal et al. (2021) worked during a very specific time period when the COVID-19 pandemic started to burst, which lead to a surge of many new trends that tried to adapt to restrictions that were caused by the deadly disease. The main objective of the research was to capture and track the temporal evolution of terms representation, using Word2Vec embeddings, in scientific papers relating to the COVID-19 pandemic, which are later fed into a machine learning system to predict emerging trends. The embeddings are trained separately for each month, which is then used to construct entities network where edges between entities are the cosine similarity between the embeddings. New links and topological features are predicted based on prior patterns of co-occurrences, thus reconnecting missing links between clinical entities over certain time intervals. With a corpus of nearly 77,000 peer-reviewed papers, the experiment had promising results, being able to capture crucial topics and findings regards to COVID-19 such as autoimmune diseases, multi-system inflammatory syndrome, and neurological complications.

2.2.3 Clustering-based Approaches

Text clustering has been a popular approach in Natural Language Processing in general, dealing with the problem of grouping text of different levels of granularity (documents, terms, paragraphs, sentences, etc.) into groups that contain similar information for ease of retrieval and browsing (Aggarwal and Zhai, 2012). In ETD, organizing text into clusters can help discover groups of popular documents/keywords based on the size of the discovered clusters. In general, three main clustering methods are commonly used for text data in ETD: K-Means clustering (MacQueen, 1967) (centroid-based), DBSCAN (Ester et al., 1996) and its extension, HDBSCAN (McInnes et al., 2017) (density-based).

Chen et al. (2007c) introduced an approach that takes advantage of a combination of statistical approach and clustering to extract hot topics. According to the paper, a term is considered "hot" if it has the following characteristics: Pervasiveness are terms that appear more often or has high Term Frequency in a time span. Topicality describes terms having positive variation in frequency through life cycle modelling using Aging Theory, a model that simulates information growth and decay (Chen et al., 2007a). With the hot terms extracted, the system generates multidimensional vectors (using exogenous features extracted from WordNet such as synonym, hypernym, as well

as endogenous hot term weight from the previous step) of sentences containing the most number of hot terms and uses Hierarchical Agglomerative Clustering (HAC) (Cutting et al., 1992) to group sentences based on similarities of their representation. Experimented on a small set of world news from March 2005 to April 2005, the empirical result showed a substantial accuracy increase in hot topic extraction compared to the classic approach using only a single vector of representation.

In an attempt to monitor mainstream and social media for firms and companies in Competitive Intelligent, a system was introduced Goorha and Ungar (2010) that targeted news articles, blog posts, review sites, and tweets to extract interesting shifts in opinions about certain products. Using word co-occurrences to identify and extract groups of phrases with a high probability of appearing next to each other, "bursty" terms with high specificity and high frequency are discovered. Moreover, using a variation of K-Means clustering for the data stream from Guha et al. (2003) with TF-IDF as features, the system was able to output quality, sensible clusters of "bursty" terms that made up topics of interest according to empirical results.

Despite being widely available digitally, digital news can be originated from multiple publishers in different languages. In order to process effectively multilingual news, an approach was proposed using sentence-BERT (Reimers and Gurevych, 2019), a siamese Transformer model that works at the sentence level, as text representation to link news articles together after monolingual batch clustering to form stories from different sources, thus creating topics with high cohesion (Linger and Hajaiej, 2020). Batch clustering is done on monolingual sets of articles published temporally close to each other using multiple TF-IDF representations of different elements of the text. After linking cluster centroids of different time periods through cosine similarity. With the distillBERT model Sanh et al. (2019), the proposed system was able to surpass previous approaches to this problem in terms of F1 and accuracy metrics in a dataset comprising English, German and Spanish documents.

In order to efficiently detect hot topics in the scenario of having an enormous amount of data to process, Li et al. (2020a) introduced a system that combines density-based clustering and parallel k-nearest neighbour topic tracking algorithm. DBSCAN is first applied in each time window to group similar text represented in TF-IDF. Afterwards, each cluster is represented by

averaging vectors of its elements and compared, in terms of cosine similarity, to news articles of a certain number of time periods. Finally, the top K most similar news articles are added to the collection for the next iteration while being considered as the label for the topic. Compared to the baseline method such as clustering using Vector Space Model with TF-IDF, the proposed system not only out-perform in terms of F1-score but also has lower time complexity, which implied high scalability.

Qiu et al. (2021) studied Personal Finance Question to track and extract new and emerging topics using corpus extracted from StackExchange[1] with labels on the topics. The approach was to encode text data at the sentence level using Sentence-Transformer and Universal Sentence Encoder (Cer et al., 2018) and then cluster the data with K-Means and HDBSCAN. The main idea was to use representation at the sentence level or reduced text representation in order to capture the occurrence of multiple themes in a single call. In addition, using representations on this level of text enables labelling the cluster by generalizing the lexical resources present in it. The task was considered a multiclass classification task and showed improvement in micro-F1 with the combination of Universal Sentence Encoder and KMeans.

2.2.4 Topic Modeling Approaches

The main objective of topic modelling is to discover word-usage patterns and associate documents of similar patterns with each other to form a topic. Topic modelling treats a topic as a probability distribution over words, a document is considered as a mixture of topics with different probability based on the words it contains (Alghamdi and Alfalqi, 2015). A popular topic modelling method used in research is LDA (Behpour et al., 2021a) which considers words representing topics has a discrete probability distribution, hence a document is regarded as a bag of words with no word order of article structure

Because of its effectiveness, LDA has been applied to numerous corpus of different specificity within the scope of ETD. Lee and Sohn (2017) aimed to identify emerging technology trends with financial business method patents by applying LDA to generate the probability of topics and assess its temporal shift, thus deciding whether a topic is in a “hot” or “cold” status. To dis-

[1] https://stackexchange.com/

tinguish the status of detected topics, survival analysis, with the combination of the Exponential Weighted Moving Average (EWMA) and Gap Time model, is utilized to determine the rise and fall in topic probability and analyze the recurrence of probability shift. It is observed that results from the proposed procedure can potentially support the Research & Development team in the research direction as it is effective to identify trends in business method patents that can signify sustainable development in the financial industry.

Behpour et al. (2021a) introduced a weighted temporal feature as a bias factor for topic clustering articles in a similar time frame. LDA is used to parameterize each abstract, followed by a Singular Value Decomposition (SVD) process to reduce the dimension of features. After K-Means clustering with a time-biased factor, to further detect trending topics, the silhouette score is divided by the standard deviation of clusters over time. The main target of the research is articles belonging to the financial sector from 1974 to 2020, where expert judgment, with a clear knowledge of the said time period, will validate the final results. Results showed that by introducing the time-biased factor in clustering, the system was able to identify interpretable topics that standard clustering could not thus increasing the trend prediction quality. Moreover, it is stated that the time-biased factor can be adjusted to suit one's needs. Lastly, the system is robust in a way that not only does it work on abstracts of journals, but it can also scale to other fields and documents of other forms.

To support experts in the decision-making process, Akrouchi et al. (2021) proposed a fully automated LDA-based system for weak signal detection. After using LDA to generate the topics within the corpus, two main filtering functions were employed. First, a "weakness" function was applied to filter topics that might contain weak signals based on three metrics: closeness centrality measures how similar one topic is to another, topic coherence within the model, and auto-correlation which tracks the relation between a number of documents per topic over a set of days and the time series. Afterwards, the "potential warning" function, which filters terms based on their normalized frequency within the topic and LDA probability, is utilized to extract early warning signs from filtered topics of the previous function. Lastly, using Word2Vec representation, words with high cosine similarity compared to terms filtered with the "potential warning" function can be considered as supports for the weak signal. Experimented on web

news from December 2019 to February 2020, the system was capable of detecting useful weak-signal related to the "epidemic", thus showing potential to support human experts in providing complementary information for decision making.

2.3 The Importance of Text Density

The length of documents in a corpus is an important factor to consider when constructing a method to analyze text since different lengths have different properties that can affect text processing in general and ETD, specifically. The amount of information encapsulated varies depending on the size of the document, thus distinct approaches can be used for the best efficiency regarding processing text. In this section, two main types of text are presented with their corresponding specific approaches and findings: sparse text that is normally seen in social media and dense text in traditional media (e.g., news articles, journals) and scientific publications.

2.3.1 Dense Text

Dense text is the type of document with a length of a few paragraphs to a few pages. News articles, journal entries, and scientific publications fall under this category. The longer text means that there is, more information within the documents, thus more latent topics to be extracted. Classic statistical approaches to dense text make use of the fact that documents of this type have a large distribution of words, hence word co-occurrences are usually the initial choice (Lin et al., 2014). However, dense text often suffers from noise and irrelevant information such as publisher information embedded in the text, articles covering not only one main subject but also a few supporting topics, or tables that could not be well-converted into text format.

To overcome such problems, a number of document structure-specific approaches have been explored. Lejeune et al. (2015) exploit the journalistic writing style of disease reports to extract information on the kind of illness and its location to detect outbreaks, targeting parts of the documents that are considered salient positions. Dridi et al. (2019), on the other hand, extract the most common keywords in titles of scientific publications and track the temporal evolution of their Word2Vec representations in the main text to explore pair of keywords with the best

similarity ranking change through time. By avoiding processing the whole text which comes with noise, these approaches have achieved some capacities of success in tracking trends in long text.

2.3.2 Sparse Text

The sparse text refers to a corpus of short texts, typically in the form of social media posts where the count of characters is limited (in Twitter's case, 256 characters). Because of the restriction in length, the content in sparse text is usually concise, thus usually containing only one topic per post (Pugachev and Burtsev, 2021). However, this sparseness hinders statistical methods that rely on word co-occurrences due to not having enough text in a single post (Lin et al., 2014). As the volume and popularity of this type of text are high (being generated every day by everyone around the world), overcoming these disadvantages is needed for the best possible extraction of a latent emerging trend.

To compensate for the lack of text in a single document, Naaman et al. (2011) took advantage of the abundant amount of posts available on Twitter to analyze potential trends in short messages. Specifically, a refined version of TF-IDF, which would favour terms appearing more often than expected, was applied to extract terms with abnormal increases in usage per hour and per day. Jin et al. (2011) tackled the aforementioned disadvantages by introducing a novel topic model called Dual Latent Dirichlet Allocation, which incorporates the dense text of a similar field to improve the consistency of extracting topics from sparse text. Another approach for detecting trends in short text introduced TF-IDF as a weighting pattern for LDA in order to capture combinations of words that are trending and not just a single token that is sometimes meaningless on its own (He et al., 2018).

2.4 Conclusions

In this chapter, we presented the state of the art for emerging trend detection in which we first discussed the use of endogenous and exogenous features in the task. Afterwards, the chapter focused on the categorization of the main approaches in the literature as followed: statistical,

graph-based, clustering, and topic modeling methods. Lastly, we introduced some specific approaches taking into account text density.

We observed that the literature has been evolving throughout the years, as researchers have found various solutions to cover the disadvantages of traditional approaches, thus complementing them. For example, term frequency evolved into TF-IDF (Jones, 2004), which was later customized with various additional weighting schemes (Fang et al., 2014), (Zhu et al., 2019). Moreover, many methods of recent years used not only one main approach but also combine with other methods to solve specific problems, creating hybrid systems that fit various scenarios in ETD (Behpour et al., 2021a; Linger and Hajaiej, 2020). With Transformer becoming a newly popular approach in various fields of natural language processing, some research has already incorporated such technology with some degree of success (Qiu et al., 2021), hence having the potential to push the state of the art further.

CHAPTER 3

Data Description

As discussed in the previous chapter, data plays a central role in emerging trend detection. Different data, in terms of time interval, text type, and sources can give specific information on trends and their variety has an influence on the methods to be designed. Thus, the discussion surrounding data cannot be neglected as it gives necessary context to the research as well as chosen methods to achieve the research objectives.

It is important to clarify first and foremost that, to the best of our knowledge, no annotated dataset is publicly available for the task of emerging trend detection, in particular in the financial domain. Thus, the variety of datasets mentioned in this chapter, among the publicly available datasets, are those that we deemed to be the most relevant to our scope of research, and we adopted them to fit the objective of the thesis. Moreover, each dataset presented in this chapter invokes a wide range of problems to solve in ETD, showing that no one-size-fits-all ETD pipeline exists to satisfy the provided data, especially in industrial settings where raw, unannotated datasets are available in large quantities.

This chapter presents four datasets in different domains that were used in our ETD research.

Firstly, we introduce DAnIEL, a dataset used in the epidemiology domain that contains news reports on the appearance of various diseases. The dataset is annotated with disease outbreaks (disease name and location) to serve the purpose of developing a system capable to identify and alert about diseases spreading.

In this experiment, we showed the differences between working with so-called clean text, and adulterated text generated from Optical Character Recognition (OCR) process. We compare these different text versions to study the impact of noise on the performance of Information Extraction systems in the same fashion as recent works examining this issue for the Named Entity Recognition task (Boros et al., 2020; Hamdi et al., 2019; Koudoro-Parfait et al., 2021) which can be a subtask of ETD. The main purpose of this study is that it is relevant, in particular for companies that wanted to have scalable systems, to discuss data quality and its effect on the performance of Information Extraction in general and Emerging Trend Detection specifically.

Next, we describe the Twitter-Trend dataset that comprises social media posts with annotations on trends. These annotations cover the trend type, to satisfy not only the task of identifying trends, but also give categories in order to determine what trends to focus on. Because of the abundance of text and annotated data in this dataset, we selected this dataset for our early study in trend discovery. While the size of textual content in Twitter-Trend is short compared to other datasets in this chapter, the uniqueness of the text lies in the fact that out-of-vocabulary words appear quite frequently (Krouska et al., 2016) and these words can denote as topics of the post, thus raises the question of removing or keeping them.

Then, provided by the LBPAM to explore potential themes and topics for future investment, we use the Event-Driven Feeds (EDF) by Bloomberg. It is a collection of news articles from many different domains such as financial, business, technology etc. While the data is enormous, the textual content in the EDF is raw. Thus, to be able to execute tasks related to ETD efficiently, cleaning noise to ensure good data quality is required. The main objective of using the EDF data is to verify whether using news articles as the main source for ETD can provide timeliness prediction.

Finally, in an entirely different domain, TRENDNert is a dataset of scientific publications that holds information on which research topics were considered as trends or lost popularity in the

Computer Science research community from 1975 to 2015. Along with the abstract text of each paper, which is usually plain text, the dataset provides knowledge on which topics the publication belongs to, as well as the movement of topics in the said time interval. Being fully annotated and having spanned across a long period, TRENDNert was selected to explore which keywords can potentially signify a trend and assess the results using provided annotation.

3.1 DAnIEL Dataset (health reports)

The corpus is dedicated to multilingual epidemic surveillance and contains articles on different press threads in the field of *health* (Google News) that focused on epidemic events from different collected documents in different languages, with events simply defined as disease-location-number of victims triplets. The corpus was built specifically for this system (Lejeune et al., 2015, 2010), containing articles from six different languages: English, French, Greek, Russian, Chinese, and Polish. It contains articles on different press threads in the field of *health* (Google News) focused on epidemic events, and it was annotated by native speakers to describe these events.

A DANIEL event is defined at document-level, meaning that an article is considered as relevant if it is annotated with a disease − location pairs (and rarely, the number of victims). An example is presented in Figure 3.1, where the event is a *listeria* outbreak in *USA* and the number of victims is unknown.

Thus, in this dataset, the event extraction task is defined as identifying articles that contain an event and the extraction of the disease name and location, i.e. the words or compound words that evoke the event. Since the events are epidemic outbreaks, there is no pre-set list of types and subtypes of events, and thus the task of event extraction is simplified to detecting whether an article contains an epidemiological event or not.

As it is rather common in event extraction, the dataset is characterized by imbalance. In this case, only around 10% of these documents are relevant (e.g. contain epidemic events), which is very sparse. The number of documents in each language is rather balanced, except for French, having about five times more documents compared to the rest of the languages. More statistics

Figure 3.1: Example of an event annotated in DANIEL dataset.


```
"15960": {
    "annotations": [
        [   "listeria",
            "USA",
            "unknown"]
    ],
    "comment": "",
    "date": "2012-01-12",
    "language": "en",
    "document_path": "doc_en/20120112_www.cnn.
        com_48eddc7c17447b70075c26a1a3b168243edcbfb28f0185",
    "url": "http://www.cnn.com/2012/01/11/health/listeria-outbreak/index.
        html"
}
```


on the corpus can be found in Table 3.1.

The DANIEL dataset is annotated at document-level, which differentiates itself from other datasets used in research for the event extraction task. A document is either reporting an event (disease-place pair, and sometimes the number of victims) or not.

Table 3.1: Summary of the DANIEL dataset. The relevant documents are documents annotated with an event.

Language	**# Documents**	**# Relevant**	**# Sentences**	**# Tokens**
French (fr)	2,733	340 (12.44%)	75,461	2,082,827
English (en)	475	31 (6.53%)	4,153	262,959
Chinese (zh)	446	16 (3.59%)	4,555	236,707
Russian (ru)	426	41 (9.62%)	6,865	133,905
Greek (el)	390	26 (6.67%)	3,743	198,077
Polish (pl)	352	30 (8.52%)	5,847	165,682
Total	4,822	489 (10.14%)	140,624	3,080,157

3.2 Twitter-Trend Typology Dataset (social media)

Zubiaga et al. (2015) created the dataset with the goal of classifying Twitter posts into trends and their respective types. In total, the dataset contains 567,452 tweets, collected from March 1st to 7th, 2011, that belong to 1,036 unique topics, thus averaging 548 tweets per topics. In addition, the dataset provides not only tweets in English as the main language (accounts for around

52.00% of total tweets), but also in 27 other languages such as Spanish (13.50%), Portuguese (11.93%), Dutch (5.58%), and Indonesian (4.03%). The topics are distributed into four classes:

- **News:** Newsworthy event that major news publishers had reported by the time the trend appeared or would report soon after it burst on Twitter. Only 142 topics (13.7%) are annotated in this category.

- **Ongoing events:** Live coverage on Twitter of a real-life event unfolding. This is the most popular class with 616 topics included (59.5%).

- **Memes:** Viral ideas or messages that are neither newsworthy nor mainstream events, but are rather funny, attractive, or commonly agreed on. 251 topics (24.2%) belong to this class.

- **Commemorative:** Celebration of anniversary (event or person). The category only contains 27 topics (2.6%)

Since the data is imbalanced regarding languages and the scope of the thesis will focus mostly on English text, tweets in other languages have been removed from the dataset. Moreover, in order to reduce imbalances in the data, trends that contain less than 10 tweets are also discarded, thus reducing the number of tweets to 338,303. Table 3.2 shows our text analysis of the English portion of the data. On average, tweets have a length of 86 characters, with a minimum length of 12 characters and a maximum length of around 140 characters, which is also the limit allowed on Twitter at the time of release of the dataset. Therefore, the dataset contains extremely short text, with only around 15 tokens per tweet in average. However, outliers existed within the data where 0.18% of tweets exceed the character limits of Twitter with unknown causes.

Table 3.2: Summary of Twitter-Trend topology dataset

	# Tweets	**# Trends**	**# Tokens (average)**	**# Characters (average)**
News	75030 (22.19%)	182 (18.70%)	14.46	84.01
Ongoing events	178797 (52.85%)	558 (57.34%)	16.44	96.05
Memes	74471 (22.01%)	204 (20.97%)	13.76	82.29
Commemorative	10095 (2.98%)	29 (2.98%)	14.58	86.59

3.3 TrendNERT Dataset (scientific articles)

The TRENDnert dataset (Moiseeva and Schütze, 2020) was released in 2017 with the purpose of creating a benchmark for uptrend and downtrend detection in scientific publications. According to the authors of TRENDnert, the dataset contains more than one million abstracts of papers on computer science published from 1975 to 2015, with which most of the entries appeared in 2000 and onward. The corpus was classified into various topics, which have been annotated to distinguish trends from downtrends. Stagnant topics that the annotators did not see any movement were not annotated and left emptied.

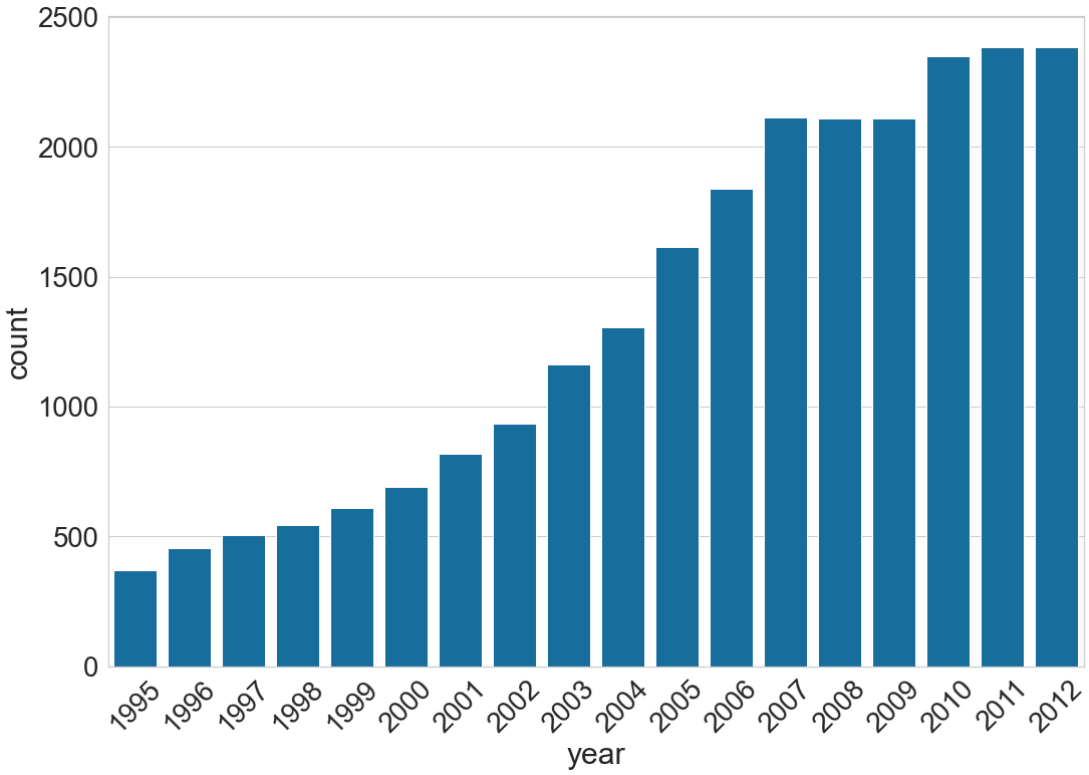


Figure 3.2: Generated TRENDnert dataset distribution from 1995 to 2012.

However, in an attempt to recreate the dataset following the guidelines given by the authors, the generated dataset was incomplete with only 155,356 entries extracted from 1995 to 2000, as presented in Figure 3.2[1], in which only more than 30 thousand entries were annotated as a (down)trend. At the end of the process, we collected 30 uptrend topics which consisted of

[1]The annotation in the dataset was not fully recreated due to a mismatch in the generated hash strings and the provided ones in the majority of entries

around 11,000 documents, and 54 downtrend topics resided in more than 20,000 papers. Table 3.3 shows the imbalance in document and topic distribution between three classes, Trend, Downtrend, and Non-relevant, where almost 80% of the generated dataset belongs to Non-relevant type and Downtrend class has almost doubled the number of topics and documents compared to Trend class.

Table 3.3: Summary of generated TRENDNert dataset

	# Topics	# Documents	# Tokens (average)
Trend	30	11,141 (7.17%)	246.84
Downtrend	54	20,626 (13.28%)	312.57
Non-relevant	N/A	123,589 (79.55%)	140.21

3.4 EDF Dataset (news articles)

In the context of this thesis, LBPAM provided us with a dataset from Bloomberg's Event-Driven Feeds (EDF). EDF data contains daily news provided by Bloomberg in the past years from 2009. Since the data is enormous (about 600 GB of raw compressed data) and it is in a deeply nested XML format with many attributes that are difficult to process, as shown in Figure 3.3, it is important to extract the raw data for query conveniences. This included parsing the raw XML data, selecting necessary information, and storing them in a database (MongoDB). At the end of the process, 6 were chosen (out of more than 20 fields), including headline, text news, date and time of publication, general topics, and tickers (stock symbols that indicate specific companies) that are considered related to the news according to Bloomberg using their proprietary algorithm. It is important to note that the EDF data does not include any annotation relating to trends, thus an approach is needed to verify and evaluate the results of trend detection, which will be discussed in later chapters.

Considering the scope of the research and the characteristic of the dataset (raw, large, and unannotated), conducting experiments and evaluating on the entire EDF data is very difficult. Thus, specific time frames and a company were chosen as follows: we studied the period from July 2019 to July 2020 - six months before the COVID-19 outbreak up until its peak, split into monthly collections of news in English. This is a very particular time span since COVID-19

triggered a massive global health crisis and drastically changed the way people live, which created new trends.

```
<ContentT EID="37183" CaptureTime="2014-08-03T00:24:23.501+00:00" Origin="
    BPOD">
     <StoryContent>
          <Id>
               <SUID>N9PFSN3MMTC0</SUID>
          </Id>
          <Event>ADD_STORY</Event>
          <Story ContentType="Current">
               <Body>
               Text Content...
               </Body>
               <BodyTextType>STYTYPE_HTML</BodyTextType>
               <Version>ORIGINAL</Version>
               <Metadata>
                    <WireId>12</WireId>
                    <ClassNum>0</ClassNum>
                    <WireName>PRN</WireName>
                    <Headline>CAIR Joins Tens of Thousands at D.C. Rally for
                        Gaza</Headline>
                    <TimeOfArrival>2014-08-03T00:24:23.235</TimeOfArrival>
               </Metadata>
               <LanguageId>1</LanguageId>
               <LanguageString>ENGLISH</LanguageString>
               <DerivedTopics>
                    <ScoredEntity>
                         <Id>Topic_Name</Id>
                         <Score>100</Score>
                    </ScoredEntity>
               <DerivedTickers>
                    <ScoredEntity>
                         <Id>Ticker_Name</Id>
                         <Score>70</Score>
                    </ScoredEntity>
               </DerivedTicker>
          </Story>
     </StoryContent>
</ContentT>
```

Figure 3.3: Example of an entry in EDF dataset.

To ease the evaluation, in absence of any ground truth trends, we wanted to work on a specific company. We chose Microsoft since it was known, at the time of this work, to play a major role in remote working trends during the pandemic by providing a stable and convenient platform for online workplace communication. Moreover, Microsoft is widely known as one of the leading companies in the field of cloud computing. Hence, it is interesting to investigate how trends in

Microsoft shifted during the pandemic. Having definite time intervals and companies that relate to a major event helps narrow down the trends to be detected, especially in a raw dataset with no annotation that contains a vast amount of information.

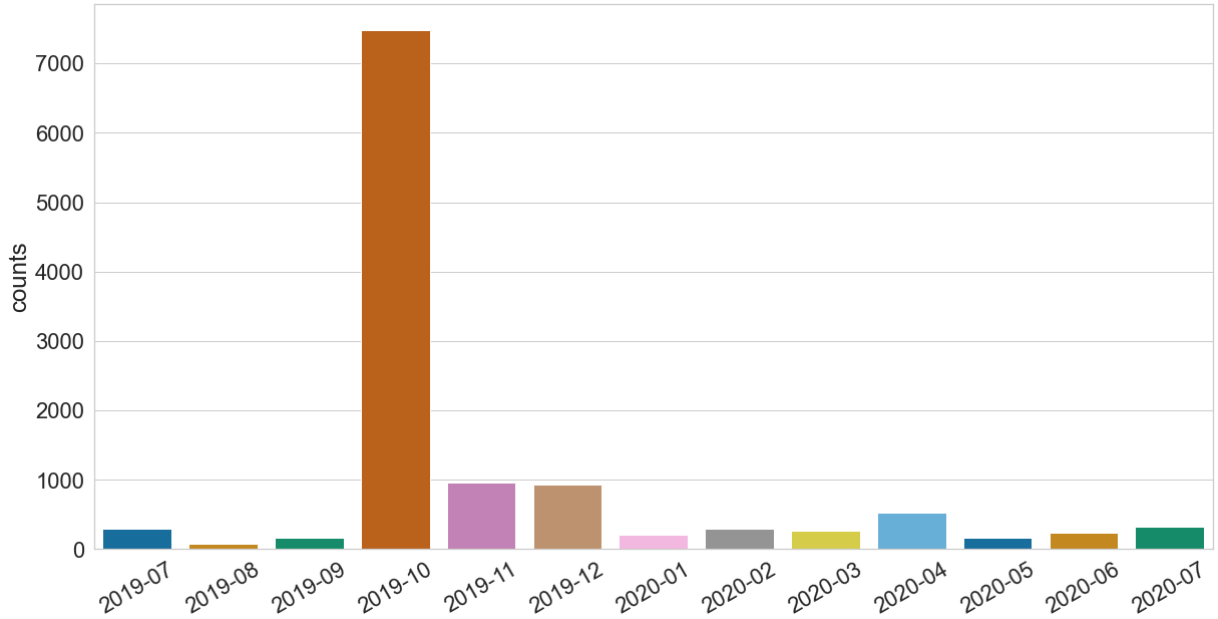


Figure 3.4: Monthly news volume distribution about Microsoft from Bloomberg EDF.

With these criteria, we extracted a total of 11,923 news articles about Microsoft within the said time span. The monthly news volume distribution is in Figure 3.4. The graph shows an enormous surge in the number of articles during the month of October 2019, which consisted of around 7500 documents and accounted for more than 60% of the total volume. In contrast, the number of documents during the beginning phase of the COVID-19 pandemic, in the span of 7 months from January 2020 to July 2020, is much lower in comparison. Averaging at about 300 articles per month, the period, the 7-month period only consisted just under 17% of the total amount of articles in the dataset. With these figures, it is expected that October 2019 could be the month that Microsoft started a trend due to the huge spike in news articles volume, while the COVID-19 period may seem uneventful to the company.

3.5 Conclusions

We described the different datasets that we used in our thesis, including various types of text, such as social media, scientific text, and news articles. The lack of annotated data specifically for the task of ETD prompted our studies and experiments to adopt and adapt to the available data in order to meet the research objectives. Moreover, the diversity in data in terms of content and annotations, or lack of it, incites different problems that cannot be solved with a single pipeline of processing.

In the next chapter, we present different studies on these datasets, based on what the dataset provides. Firstly, we experiment with how noise can impact the task of information extraction in a research surrounding the DAnIEL dataset and the corresponding epidemic surveillance system. Afterward, we describe the results of different pre-processing strategies on our dataset, having impact on tasks according to annotation availability.

CHAPTER 4

Impact of Data Quality and Text Pre-processing on ETD and other tasks

In order to effectively detect emerging trends, distinct pre-processing strategies need to be considered for each set of data to remove unwanted noise and redundant information, thus retaining only the main content of the document that potentially provides useful knowledge for ETD methods to extract from. For unstructured and semi-structured data, such as the data we presented in Chapter 3, pre-processing is deemed necessary due to them retaining the whole text without any modification. Undesirable text that can skew the result of ETD and needs to be removed using processing may include, but is not limited to, stopwords (Saif et al., 2014), text boilerplate (Barbaresi and Lejeune, 2020)[1]), duplicated content, such as highly repeated phrases (Suchomel, Pomikálek, et al., 2012), etc. Moreover, when converting articles to text format, tables are usually not interpreted properly, hence do not appear as tables in the text and are not marked as tables, making the section of the text confusing and difficult to process. Because of

[1]Text boilerplate is the portion of the article that exist in every article but contains no useful content (for example, publisher's information, detail about the author

this difficulty, converted tables usually cause noise and are unnecessary to the main content.

In this chapter, the main objective is to explore the impact of data quality and the importance of preparing the data, and pre-processing, before executing tasks related to Information Extraction in general and ETD specifically. Often, it is said that the text pre-processing phase is done at the beginning of every task, yet its influence with respect to different datasets and different tasks is not clearly explained. Because of this, we are interested in discovering how impactful data quality is and whether having text pre-processed is as mandatory as suggested in various tutorials, guides, and research papers.

We, first, give our insights regarding the impact of noise on the result of Information Extraction by presenting a case study with epidemic outbreak surveillance by exploring the DAnIEL system (Brixtel et al., 2013) and the annotated dataset that comes along with the system (Mutuvi et al., 2020). This system has the main purpose of identifying the spread of diseases to obtain necessary information for health authorities to respond accordingly (Lejeune et al., 2015). Since the objective of DAnIEL aligns closely with emerging trend detection, the study is included as part of the research to explore the importance of data cleaning and pre-processing. Next, we present specific pre-processing approaches to remove noise and redundancy existing in the datasets presented in Chapter 3. The main focus of this study is on the Twitter-Trend and EDF datasets, as the text from each of these corpora required different noise removal directions. In the case of Twitter-Trend, one should consider the removal of special characters since they are quite frequent, with some being part of important words/terms. As for the EDF data, its raw text, containing many difficult texts to process such as boilerplate and tables, required full attention to determining the pre-processing strategy. As for the TRENDnert dataset, after analyzing the corpus, we observed that this dataset exhibits different characteristics: the text content is simple with only alphabetic and numeric characters without special characters. Therefore, we considered that having only stopwords removed is sufficient. Thus, this point will not be discussed any further in this chapter.

The experiments in this chapter have resulted in the following publications: "*Impact Analysis of Document Digitization on Event Extraction*" (Nguyen et al., 2020), and "*Assessing the impact of OCR noise on multilingual event detection over digitised documents*" (Boros et al., 2022).

4.1 Impact of Data Quality

The first part of this chapter investigates how data quality can affect the performance of Information Extraction systems. We selected the scenario of epidemic outbreak surveillance through text for this experiment for numerous reasons. First and foremost, disease outbursts can be considered a type of trend in the context of health emergencies, causing negative effects in different aspects of daily life while forcing official authorities to act in response to the threat of the outbreak for the safety of civilians. Hence, detecting epidemic outbreak is not only beneficial in terms of health care but also identify potential impact on the economy and society. Secondly, as we have mentioned in previous chapters, within the context of ETD, we lack the necessary annotated datasets to realize our experiments. However, in the case of epidemic surveillance, annotated data is available, thus allowing us to fully conduct evaluations on the scenarios we created.

4.1.1 In the Context of Health Reports (DAnIEL)

The surveillance of epidemic outbreaks has been an ongoing challenge globally, and it has been a key component of public health strategy to contain diseases spreading. While digital documents have been the standard format in modern days, many archives and libraries still keep printed historical documents and records. Historians and geographers have a growing interest in these documents as they still hold much crucial information and events in the past to analyze, noticeably in health and related to epidemics events in an international context.

However, due to the digitization process, several issues can arise, most common is the case when the original document is distorted, whether through deterioration due to aging or was damaged in the storing process, which will affect the converted content. Moreover, errors from the digitization process could also be a factor that causes adulteration of the converted documents, e.g. word variations or misspellings, in other words, noise (Goodman et al., 2016).

DAnIEL (Lejeune et al., 2015) stands for Data Analysis for Information Extraction in any Language. The approach is at the document-level, as opposed to the commonly used analysis at the sentence-level, by exploiting the global structure of news as defined by Lucas (2009). The

entries of the system are news texts, the title and body of text, the name of the source when available, and other metadata (e.g date of article). As the name implies, the system has the capability to work in a multilingual setting due to the fact that it is not a word-based algorithm, segmentation in words can be highly language-specific, but rather a character-based one that centres around the repetition and position of character sequences.

By avoiding grammar analysis and the usage of other NLP toolkits (e.g part-of-speech tagger, dependency parser) and by focusing on the general structure of journalistic writing style (Hamborg et al., 2018b; Lucas, 2009), the system is designed to ease a multilingual coverage. The system is able to detect crucial information in salient zones that are peculiar to this genre of writing: the properties of the journalistic genre and the style universals form the basis of the analysis. This combines with the fact that DANIEL considers text as sequences of characters, instead of words, the system can quickly operate on any foreign language and extract crucial information early on and improve the decision-making process. This is pivotal in epidemic surveillance since timeliness is key, and more than often, initial reports where patient zero appears are in the vernacular language. The approach presented by Lejeune et al. (2015) considers the document as the main unit and aims at the language-independent organizational properties that repeat information at explicit locations. According to the author, epidemic news reports, which use the journalistic writing style, have well-defined rules on structure and vocabulary to convey concisely and precisely the message to their targeted audience. As these rules are at a higher level than grammar rules conceptually, they are applied to many languages, thus offering high robustness in a multilingual scenario.

We aim to test this model's robustness against noise, in other words, its ability to treat highly inflected languages and misspelt or unseen words, which can be either due to the low quality of text or the spelling variants. We compare it to a more classical Machine Learning based approach. For these experiments, we present the general evaluation settings. Furthermore, we create synthetic data starting from the initial dataset in order to study the direct impact of automatic text recognition (ATR) over the performance of both approaches.

The focal point of this set of experiments is to observe how the level of noise stemming from the digitization process impacts the performance of the models. However, there is no adequate

historical document dataset provided with manually curated event annotation that could directly be used to measure the performance of the models over deteriorated historical documents. Thus, the noise and degradation levels have to be artificially generated from clean documents, to measure the impact of ATR over event detection using DAnIEL. We shall thus use readily available data sets over contemporary and digitally-born datasets, which are free of any ATR-induced noise.

In order to create such an appropriate dataset, the raw text from the DANIEL dataset was extracted and converted into clean images. For simulating different levels of degradation, we used DocCreator (Journet et al., 2017). The rationale is to simulate what can be found in deteriorated documents due to time effects, poor printing materials, or inaccurate scanning processes, which are common conditions in historical newspapers. We used four types of noise:

Character Degradation : adds small ink dots on characters to emulate the age effect on articles;

Phantom Character : simulates when characters appearance erodes due to excessive use of documents;

Bleed Through : bleeding through appears in double-paged document image scans where the content of the back side appears in the front side as interference;

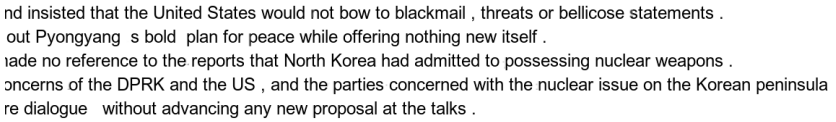
Blur : blurring is a common degradation effect encountered during a typical digitization process.

After contaminating the corpus, all the text was extracted from noisy images[2], for initial clean images (without any adulteration) and the noisy synthetic ones. An example of the degradation levels is illustrated in Figure 4.1. The noise levels were empirically chosen with a considerable level of difficulty[3].

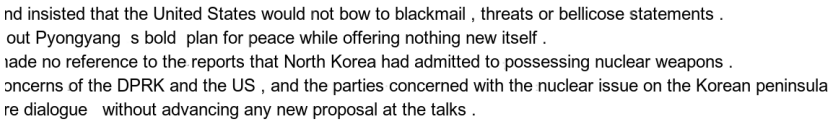
The experiments were conducted in the following manner: for each noise type, a different intensity is generated to see its relation to the performance of the model. Character error rate

[2]The Tesseract optical character recognition (OCR) Engine v4.0 https://github.com/tesseract-ocr/tesseract (Smith, 2007) was used to produce the digitised documents.

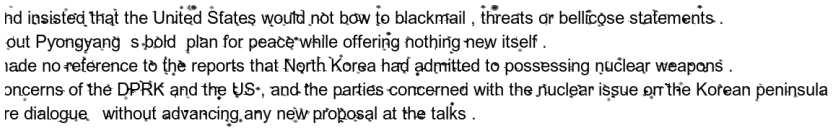
[3]The following values of DocCreator are: *Character Degradation* (2-6), *Phantom Character* (Very Frequent), *Blur* (1-3), *Bleed Through* (80-80).



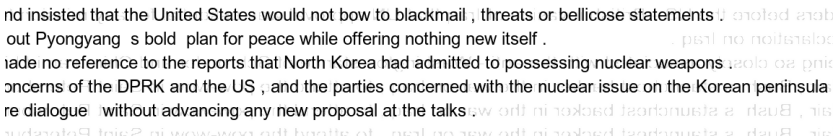
(a) Original (clean) image



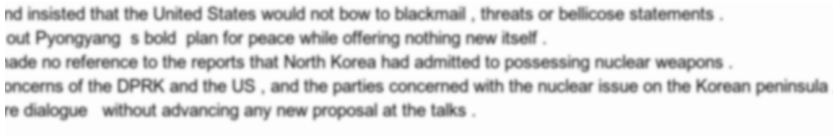
(b) Phantom Character



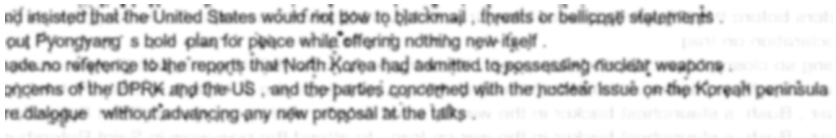
(c) Character Degradation



(d) Bleed Through



(e) Blur



(f) All types of noises applied together.

Figure 4.1: Original image and result of the different levels of noise to the image.

(CER) and word error rate (WER) were calculated for each noise level. These two measures can align long noisy text even with additional or missing text with the ground truth, thus making it possible to calculate the error rate of the OCR process. The experiments are performed under conditions of varying word error rate (WER) and character error rate (CER): original text, OCR from high-quality text images, and OCR on synthetically degraded text images.

4.1.1.1 Evaluation Framework

For the evaluation of the performance of the event detection task, we use the standard metrics: Precision (P), Recall (R), and F-score (F1). Precision measures the rate of correct predictions (True Positives) made by the system with respect to all the positive decisions (True and False Positives). Recall evaluates the percentage of true positives predicted with respect to all the positive instances of the dataset (True Positives and False Negatives). Lastly, F-score is a harmonic mean computed from precision and recall values in order to give an overall view of the performance of the system. For measuring the document distortion due to the OCR process, we also report the standard metrics: CER and WER. We perform two types of evaluations, both at the document level (which corresponds to the level of annotation in the DANIEL dataset):

- Event identification: a document represents an event if both triggers were found, regardless of their types;
- Event classification: a document represents an event if the triggers are correctly found and match exactly with the ground truth ones.

4.1.2 Experiments with Clean DAnIEL dataset

Hereafter, we present the experiments performed with the clean data. Considering that the DANIEL system has a ratio parameter for matching the extracted triggers, we test two values for it. For the first experiments, we use a ratio value of 0.8 (the default value of the system) that was empirically chosen by Lejeune et al. (2015) for the best trade-off between recall and precision. Second, we test the maximum ratio value of 1.0 in order to analyze the system's performance when the extracted disease names and locations exactly match with the knowledge base.

Table 4.1: Evaluation of DANIEL on the initial dataset for event identification (regardless of the types of the triggers).

		Polish	Chinese	Russian	Greek	French	English	All languages
ratio=0.8	P	0.6842	0.8	0.7115	0.641	0.592	0.4918	0.6052
	R	0.8667	1.0	0.9024	0.9259	0.9088	0.8571	0.9059
	F1	0.7647	0.8889	0.7957	0.7576	0.7169	0.625	0.7256
ratio=1.0	P	0.0	0.0	0.0	0.0	0.9155	0.0	0.9155
	R	0.0	0.0	0.0	0.0	0.5735	0.0	0.3988
	F1	0.0	0.0	0.0	0.0	0.7052	0.0	0.5556

For event identification on clean textual data, one can notice from Table 4.1, that usually DANIEL favors recall instead of precision and tends to suffer from an imbalance between precision and recall, which may be due to the high imbalance of the data. It is also not surprising that the DANIEL system has the highest performance values for event identification for Chinese and Greek, since for Chinese, there are few relevant documents compared with the other languages (16 documents that report an event), and for Greek, there are 26 of them.

We also can note the large difference between the two chosen ratios. More exactly, an increase in this value comes in the detriment of the languages that are not only morphologically rich but also in the case where the exact name of the disease is not located in the text.

Table 4.2: Evaluation of DANIEL for event classification (triggers are correctly found and match with the ground truth ones).

		Polish	Chinese	Russian	Greek	French	English	All languages
ratio=0.8	P	0.3421	0.35	0.2692	0.4103	0.5211	0.2951	0.4645
	R	0.4	0.4118	0.3146	0.5079	0.5781	0.4737	0.5363
	F1	0.3688	0.3784	0.2902	0.4539	0.5481	0.3636	0.4978
ratio=1.0	P	0.0	0.0	0.0	0.0	0.7934	0.0	0.7934
	R	0.0	0.0	0.0	0.0	0.3592	0.0	0.2666
	F1	0.0	0.0	0.0	0.0	0.4945	0.0	0.3991

In the case of event classification, we observe from Table 4.2, that DANIEL is balanced regarding the precision and recall metrics, being able to have higher F1 on the under-represented languages (Chinese, Russian, and Greek). We also notice that, in all the cases, DANIEL does not detect the number of victims. We assume that this is due to the fact that many of the annotated numbers cannot be found in the text, e.g. $10,000$ cannot be detected since the original

text has the $10,000$ form, or it is spelled *ten thousand*. Generally, for the detection of locations, we recall that DAnIEL is capable to detect locations due to the usage of external resources and article metadata.

For the experiments on noisy data, we will use a ratio value of 0.8, since the maximum value for the ratio creates results prone to suffering from word variations or word misspellings (which is a direct consequence of the digitization process).

4.1.3 Experiments with Noisy DAnIEL dataset

The results in Table 4.3 clearly state that *Character Degradation* is the effect that affects the transcription of the documents the most. However, for character-based languages (e.g. Chinese), CER is commonly used instead of WER as the measure for OCR, and, thus, we report only the CER (Wang et al., 2013).

Table 4.3: Document degradation OCR evaluation on the DANIEL dataset.

		Clean	CharDeg	Bleed	Blur	Phantom	All
All	CER	2.61	9.55	2.83	8.76	2.65	11.07
	WER	4.23	26.23	5.93	19.05	4.71	27.36
Polish	CER	0.15	5.86	0.19	7.57	0.19	5.51
	WER	0.74	20.66	1.17	13.23	1.17	20.70
Chinese	CER	36.89	41.01	38.24	43.97	36.91	46.97
	WER	–	–	–	–	–	–
Russian	CER	0.93	16.20	1.45	8.13	1.03	10.91
	WER	1.63	28.46	6.61	14.94	2.73	29.72
Greek	CER	3.52	9.04	3.76	13.79	3.54	16.28
	WER	15.86	41.36	17.39	54.02	15.93	54.76
French	CER	1.96	8.37	2.13	7.43	2.0	10.90
	WER	3.33	23.56	4.89	16.31	3.76	26.07
English	CER	0.35	5.75	0.52	4.74	0.44	7.43
	WER	0.66	24.78	2.14	14.72	1.66	20.99

We note also that, regarding the Chinese documents, the high values for CER, for every type of noise, might be caused by the existence of the enormous number of characters in the alphabet that, by adding such an effect as *Character Degradation* can change drastically the recognition of a character (and in Chinese, one single character can often be a word). Otherwise, while

Character Degradation noise and *Blur* effect have more impact on the performance of DANIEL than *Phantom Character* type since it did not generate enough distortion to the images. A similar case applies to the *Bleed Through* noise.

Regarding the experiments presented in Tables 4.4 and 4.5, we notice, first of all, that the *Character Degradation* effect, *Blur*, and most of all, all the effects mixed together, have indeed an impact or effect over the performance of DAnIEL, but with little variability. Meanwhile, *Phantom Degradation* and *Bleed through* had very little to no impact on the quality of detection with DAnIEL.

Table 4.4: Evaluation of DANIEL results on the noisy data for event identification (regardless of the types of the triggers). Orig=Original, PL=Polish, ZH=Chinese, RU=Russian, EL=Greek, FR=French, EN=English.

		Orig	Clean	CharDeg	Bleed	Blur	Phantom	All
All	P	0.61	0.735 (+0.12)	0.755 (+0.14)	0.735 (+0.12)	0.74 (+0.13)	0.731 (+0.12)	0.758 (+0.14)
	R	0.91	0.859 (-0.05)	0.674 (-0.23)	0.862 (-0.04)	0.857 (-0.05)	0.862 (-0.04)	0.718 (-0.19)
	F1	0.73	0.792 (+0.06)	0.712 (-0.01)	0.793 (+0.06)	0.794 (+0.06)	0.791 (+0.06)	0.737 (+0.00)
PL	P	0.68	0.643 (-0.03)	0.656 (-0.02)	0.658 (-0.02)	0.692 (+0.01)	0.643 (-0.03)	0.645 (-0.03)
	R	0.87	0.9 (+0.03)	0.7 (-0.17)	0.9 (+0.03)	0.9 (+0.03)	0.9 (+0.03)	0.667 (-0.20)
	F1	0.76	0.75 (-0.01)	0.677 (-0.08)	0.761 (+0.00)	0.783 (+0.02)	0.75 (-0.01)	0.656 (-0.10)
ZH	P	0.8	0.882 (+0.08)	0.882 (+0.08)	0.789 (-0.01)	0.733 (-0.06)	0.789 (-0.01)	0.857 (+0.05)
	R	1.0	0.938 (-0.06)	0.938 (-0.06)	0.938 (-0.06)	0.917 (-0.08)	0.938 (-0.06)	0.75 (-0.25)
	F1	0.89	0.909 (+0.01)	0.909 (+0.01)	0.857 (-0.03)	0.815 (-0.07)	0.857 (-0.03)	0.8 (-0.09)
RU	P	0.71	0.688 (-0.02)	0.691 (-0.01)	0.688 (-0.02)	0.705 (-0.00)	0.688 (-0.02)	0.727 (+0.01)
	R	0.9	0.805 (-0.09)	0.744 (-0.15)	0.846 (-0.05)	0.795 (-0.10)	0.846 (-0.05)	0.821 (-0.08)
	F1	0.8	0.742 (-0.05)	0.716 (-0.08)	0.759 (-0.04)	0.747 (-0.05)	0.759 (-0.04)	0.771 (-0.02)
EL	P	0.64	0.59 (-0.05)	0.682 (+0.04)	0.59 (-0.05)	0.639 (-0.00)	0.59 (-0.05)	0.667 (+0.02)
	R	0.93	0.852 (-0.07)	0.556 (-0.37)	0.852 (-0.07)	0.852 (-0.07)	0.852 (-0.07)	0.518 (-0.41)
	F1	0.76	0.697 (-0.06)	0.612 (-0.14)	0.697 (-0.06)	0.73 (-0.03)	0.697 (-0.06)	0.583 (-0.17)
FR	P	0.59	0.803 (+0.21)	0.828 (+0.23)	0.806 (+0.21)	0.801 (+0.21)	0.801 (+0.21)	0.816 (+0.22)
	R	0.91	0.849 (-0.06)	0.666 (-0.24)	0.849 (-0.06)	0.849 (-0.06)	0.849 (-0.06)	0.723 (-0.18)
	F1	0.72	0.826 (+0.10)	0.738 (+0.01)	0.827 (+0.10)	0.825 (+0.10)	0.825 (+0.10)	0.767 (+0.04)
EN	P	0.49	0.508 (+0.01)	0.458 (-0.03)	0.508 (+0.01)	0.516 (+0.02)	0.508 (+0.01)	0.52 (+0.03)
	R	0.86	0.943 (+0.08)	0.629 (-0.23)	0.943 (+0.08)	0.943 (+0.08)	0.943 (+0.08)	0.743 (-0.11)
	F1	0.62	0.66 (+0.04)	0.53 (-0.09)	0.66 (+0.04)	0.667 (+0.04)	0.66 (+0.04)	0.612 (-0.00)

The cause of the decrease in performance of DANIEL is that, in order to detect events, the system looks for repeated substrings at salient zones. In the case of many incorrectly recognized

words during the OCR process, there may be no repetition anymore, implying that the event will not be detected. However, since DANIEL only needs two occurrences of its clues (substring of a disease name and substring of a location), it is assumed to be robust to the loss of many repetitions, as long as two repetitions remain in salient zones.

Table 4.5: Evaluation of DANIEL results on the noisy data for event classification (triggers are correctly found and match with the ground truth ones). Orig=Original, PL=Polish, ZH=Chinese, RU=Russian, EL=Greek, FR=French, EN=English.

		Orig	Clean	CharDeg	Bleed	Blur	Phantom	All
All	P	0.46	0.552 (+0.09)	0.548 (+0.08)	0.549 (+0.08)	0.558 (+0.09)	0.548 (+0.08)	0.547 (+0.08)
	R	0.54	0.497 (-0.04)	0.377 (-0.16)	0.496 (-0.04)	0.497 (-0.04)	0.498 (-0.04)	0.4 (-0.14)
	F1	0.5	0.523 (+0.02)	0.447 (-0.05)	0.521 (+0.02)	0.526 (+0.02)	0.521 (+0.02)	0.462 (-0.03)
PL	P	0.34	0.333 (-0.00)	0.328 (-0.01)	0.342 (+0.00)	0.359 (+0.01)	0.333 (-0.00)	0.274 (-0.06)
	R	0.4	0.431 (+0.03)	0.323 (-0.07)	0.431 (+0.03)	0.431 (+0.03)	0.431 (+0.03)	0.262 (-0.13)
	F1	0.37	0.376 (+0.00)	0.326 (-0.04)	0.381 (+0.01)	0.392 (+0.02)	0.376 (+0.00)	0.268 (-0.10)
ZH	P	0.35	0.412 (+0.06)	0.353 (+0.00)	0.342 (-0.00)	0.367 (+0.01)	0.342 (-0.00)	0.464 (+0.11)
	R	0.41	0.412 (+0.00)	0.353 (-0.05)	0.382 (-0.02)	0.423 (+0.01)	0.382 (-0.02)	0.382 (-0.02)
	F1	0.38	0.412 (+0.03)	0.353 (-0.02)	0.361 (-0.01)	0.393 (+0.01)	0.361 (-0.01)	0.419 (+0.03)
RU	P	0.27	0.302 (+0.03)	0.312 (+0.04)	0.302 (+0.03)	0.295 (+0.02)	0.302 (+0.03)	0.273 (+0.00)
	R	0.31	0.326 (+0.01)	0.357 (+0.04)	0.341 (+0.03)	0.306 (-0.00)	0.341 (+0.03)	0.282 (-0.02)
	F1	0.31	0.314 (+0.00)	0.333 (+0.02)	0.32 (+0.01)	0.301 (-0.00)	0.32 (+0.01)	0.278 (-0.03)
EL	P	0.41	0.333 (-0.07)	0.341 (-0.06)	0.333 (-0.07)	0.361 (-0.04)	0.333 (-0.07)	0.357 (-0.05)
	R	0.51	0.413 (-0.09)	0.238 (-0.27)	0.413 (-0.09)	0.413 (-0.09)	0.413 (-0.09)	0.238 (-0.27)
	F1	0.45	0.369 (-0.08)	0.28 (-0.17)	0.369 (-0.08)	0.385 (-0.06)	0.369 (-0.08)	0.286 (-0.16)
FR	P	0.47	0.691 (+0.22)	0.693 (+0.22)	0.69 (+0.22)	0.689 (+0.21)	0.689 (+0.21)	0.675 (+0.20)
	R	0.51	0.527 (+0.01)	0.402 (-0.10)	0.524 (+0.01)	0.527 (+0.01)	0.527 (+0.01)	0.431 (-0.07)
	F1	0.49	0.598 (+0.10)	0.509 (+0.01)	0.596 (+0.10)	0.597 (+0.10)	0.597 (+0.10)	0.526 (+0.03)
EN	P	0.47	0.292 (-0.17)	0.26 (-0.21)	0.292 (-0.17)	0.297 (-0.17)	0.292 (-0.17)	0.31 (-0.16)
	R	0.51	0.5 (-0.01)	0.329 (-0.18)	0.5 (-0.01)	0.5 (-0.01)	0.5 (-0.01)	0.408 (-0.10)
	F1	0.49	0.369 (-0.12)	0.291 (-0.19)	0.369 (-0.12)	0.372 (-0.11)	0.369 (-0.12)	0.352 (-0.13)

Regarding all the aforementioned results for the DANIEL system, computing the number of affected event words (disease, location, number of cases), we also notice that a very small number of them have been modified by the OCR process, only 1.98% for all the languages together, for all the effects mixed together, close to the 1.63% that were affected by the OCR on clean data. This is due to the imbalance in the DANIEL dataset: only 10.14% of a total of $4,822$ documents contain events. It brings us to the conclusion that the event extraction task is not considerably impacted by the degradation of the image documents.

One interesting observation is that the precision or recall can increase, resulting in a higher F1, despite the higher noise effect applied. One possible explanation for this phenomenon is that with a greater level of noise, some false positives disappear. Documents, which were previously classified wrongly due to being too ambiguous to the system (for instance documents related to vaccination campaigns are usually tagged as non-relevant in the ground truth dataset), were given much more distinction thanks to the noise, thus making them look less like relevant samples to the system. More formally: let document X be a false positive in its raw format (X_{raw}). Let X_{Noisy} be its noisy version. If the paragraph that triggered both systems' misclassifications disappeared in X_{noisy}, there is a good chance that it will be classified as non-relevant. In that case, X_{raw} is a false positive but X_{noisy} is a true negative. That may seem counter-intuitive, but noise can improve classification results, see for instance (Lejeune and Zhu, 2018) for a study on the same dataset of the influence of boilerplate removal on results.

4.1.4 Conclusions and Discussion

We conclude that, in our experimental setting, while the DANIEL system is a robust solution with respect to noisy data by not having its performance reduced considerably, the epidemic information extraction task is definitely prone to noise. Moreover, while these experiments were performed in an artificial setting with synthetically produced noise effects, a more realistic scenario could generate other tremendous issues due to the digitization process. Nevertheless, the study paves the way for future work on ETD in printed news articles as text is not the only source of newspapers.

4.2 Impact of Text Pre-processing

This section dedicates to our testing with different text pre-processing procedures with the Twitter-Trend and Event-Driven Feed data. As established in the case study with DAnIEL in section 4.1, while the method can be prone to noise, it can still have good performance despite the low data quality. Because of this, we are intrigued to examine how text pre-processing can impact the results of different tasks dedicated to the data and whether it is mandatory to perform this phase beforehand. As mentioned previously, we purposely exclude the TRENDNert dataset

due to having a simple writing format with no noteworthy element to be removed.

4.2.1 In the Context of Social Media (Twitter-Trend)

At the pre-processing phase of the data, due to its extremely limited length which leads to the text being quite simple with one to two sentences in each tweet expressing opinions of users, the initial strategy is quite straightforward with the inclusion of only stopwords removal and text standardization with character lowercasing. Nonetheless, Twitter data usually contains many special characters per post, as shown in Figure 4.2, to tag an user, include a hashtag that follows certain events/people or expresses emotion through emoticons. Because of this, it should be taken into consideration if removing special characters is necessary to avoid additional noise in the dataset.

```
@twitter_username did I just mention Bon Jovi?
March 2nd is Jon Bon Jovi's birthday! #bonjovi
I just got this tweet from them, haha! Happy bday Jon <3
```

Figure 4.2: A sample Twitter post (anonymyzed)

In order to decide whether special characters are unwanted text to be disregarded, we experimented with removing non-alphabet characters to identify their effect on classifying text into trends. Two classic methods of classification were used, Linear SVM and Multinomial Naive Bayes with the feature used in both methods being TF-IDF. This bag-of-words feature was chosen with the purpose of distinguishing words with and without special characters, thus skewing the distribution of words in order to observe its effect on the performance of classification. The dataset is split into two parts, 75% for the train set and the remaining 25% for the test set. To evaluate the results, we again use the standard metrics for classification evaluation: precision, recall and f-score.

Results from Table 4.6 show that removing special characters during the pre-processing phase improves the performance of both methods. While the differences in evaluation metrics between the two pre-processing strategies were marginally close in the case of classification using Linear SVM, the distinction was more apparent when classifying with Multinomial Naive Bayes, most noticeable in recall value. From this experiment, we conclude that discarding special characters

should be done in order to achieve the cleanest text, as even in a dataset where special characters are a crucial part of the text, the performance in classification is still better when having them removed.

Table 4.6: Comparing the result of twitter trend classification with different text pre-processing strategy

		Precision	Recall	F-Score
Linear SVM	Remove special characters	0.94	0.94	0.94
	Keep special characters	0.93	0.92	0.92
Multinomial Naive Bayes	Remove special characters	0.86	0.85	0.84
	Keep special characters	0.82	0.79	0.77

4.2.2 In the Context of News Articles (EDF)

As mentioned previously in Chapter 3, only the selected portion of the data about Microsoft from July 2019 to July 2020 will be considered. From our initial observation of the raw data, plenty of duplicate entries exist within the dataset, since the dump provided by Bloomberg contains not only the news but also any updated version of them, however slight they are. Moreover, since the main interest is the textual content, stock market reports containing tickers and a large number of numbers in them do not concern us. Thus, it is crucial to filter these unnecessary documents in order to be certain that the quality of the final structured dataset is prepared for further investigation.

From our analysis of the data, there are text patterns that appear relatively frequent in the corpus and mainly talk about Bloomberg's own publisher's information, more exactly, Bloomberg's standardized text containing contact address and phone number, as shown in Figure 4.3. This provides no valuable knowledge to our task. On the contrary, this standardized text could actually hinder the performance of a system as it could cause potential keywords to be more subtle, thus harder to be recognized. Therefore, we remove all of Bloomberg's standardized text as part of the data-cleaning process.

Moreover, Bloomberg News is geared towards an audience that is interested in finance, thus articles are written in a way that usually contains an abundance of words that belong to the financial glossary (e.g. "dividend", "bond", "equity"). Because the goal is to look for topics that

```
--Editor: XXX
To contact the reporter on this story:
XXX in New York at +1-XXX-XXX-XXXX or
XXX@bloomberg.net
To contact the editor responsible for
this story:
XXX at +1-XXX-XXX-XXXX or
XXX@bloomberg.net
View source version on businesswire.com:
http://www.businesswire.com/news/home
/2016040XXX5296/en/
Contact:
XXX@bloomberg.net
-0- Apr/07/2016 12:00 GMT
```

Figure 4.3: An example of Bloomberg standardized text (anonymized).

do not specifically belong to the financial sector, the existence of this vocabulary could cause the same problem as with Bloomberg's standardized text, thus is it also removed from the text. Afterward, we perform a lemmatization to return to the canonical form of the token in the text, for reducing the size of the vocabulary to process in later steps. Lastly, any article that has less than ten tokens is considered uninformative and is removed[4].

In addition to making the removal of boilerplate and financial vocabulary an obligatory part of the pre-processing, we also consider special characters, stopwords, and numbers to be discarded. Specifically, in the case of numeric characters, we discovered that due to the reporting nature of news articles, they can contain tables to present data, which are not converted properly to text format nor have any marking to signify a table. Thus, in many articles, there exist a high influx of numeric characters. With this observation, we intend to experiment with whether it is necessary to keep all the numeric characters or they can be removed, not only to keep the data quality high with less unusable text but also to significantly reduce the amount of text to process and save some computational time. Consequently, removing special characters, numbers, and stopwords is experimented with separately along with discarding them altogether. Furthermore, we also compare these results to that of the raw text with none of the elements above removed

[4]For this step, we manually checked different values of this threshold and the minimum quality of the remaining per-processed article.

to evaluate the effectiveness of pre-processing step.

As mentioned in Chapter 3, the EDF dataset is a raw, unannotated dataset, thus requiring us to adopt unsupervised methods. Hence, the task we experimented on to assess the impact of every pre-processing scheme is Latent Dirichlet Allocation, which is a topic modeling method capable of generating interpretable topics stated in Chapter 2. Since it is required to know the number of topics in advance of using LDA and given the size of the data we extracted from the EDF in section 3.4, the number of the topics chosen for this experiment starts from 20 to 80 with the step count of two, to ensure that the experiment can cover enough ranges of topics. Consequently, text from each type of pre-processing will go through 41 LDA models to generate topics for assessment.

In order to evaluate the results of LDA in each data cleaning strategy, we utilized the Coherence Score metric to gauge how correlated the deducted topics are to one another. According to a study by Röder et al. (2015), it is discovered that a measure labeled C_V is the best for assessing topic coherence, scoring the highest correlation with all available human rating data in the research. Therefore, the C_V metric is adopted for topic coherence calculation in our experiment.

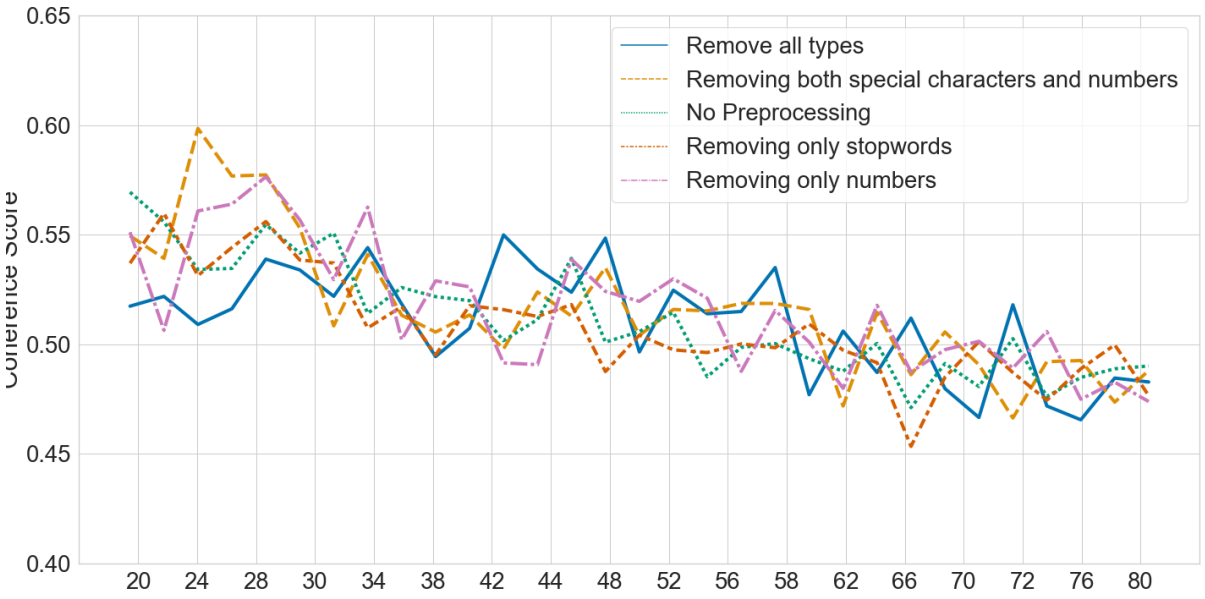


Figure 4.4: Comparing LDA coherence score on different pre-process strategy on the EDF data on Microsoft from July 2019 to July 2020

To calculate C_V coherence, four phases are executed: (1) Firstly, top-N words of each resulting topic are segmented into pairs in order to later measure the extent of how each word supports or undermines the other. (2) Afterward, the probability of a single word or a pair of words is calculated using Boolean document calculation, which takes the number of documents a word/pair of words appears in and divided it by the total documents count, thus ignoring the total frequencies of words and the distance between words in pair (how far they are situated from each other in the document). In order to capture the proximity of words, a Boolean sliding window is implemented to consider a set number of tokens as a virtual document in a step and each step moving forward creates an entirely new virtual document where every virtual document applies the Boolean document calculation to calculate word probability. (3) A confirmation measure is then calculated for each segmentation done in the previous step, taking into account their probability, by first calculating the normalized pointwise mutual information (NPMI) of each pair of words in the segmentation to formulate a context vector. With the context vectors calculated for each segmentation, cosine similarity between each pair of vectors is used to calculate the confirmation measure. (4) Lastly, the coherence score is aggregated by taking the mean of all confirmation measures.

Figure 4.4 shows the result of our experiment, illustrating the change in coherence score for each type of pre-processing as the number of topics increases. It is important to note that the result of removing only numbers and only special characters is identical to one another, thus implying that they have a similar impact to LDA. Thus, we decided to include another type of pre-processing in our experiment that removes both special characters and numbers. Surprisingly, none of the stated strategies has any outstanding improvement compared to one another or even with no pre-processing. In some cases with a specific number of topics, one type of pre-processing is better than the other. For example, the highest coherence score is when both special characters and numbers are removed, where the number of topics is 28. However, it is the only instance where the result of this type of pre-processing outperforms noticeably compared to others. When combining the removal of stopwords, numbers, and special characters, the generated LDA topics have the highest coherence score with clear distinction in topic numbers of 42, 48, 58, and 72. Interestingly, the text produced from no pre-processing has the highest result when the number of topics is 20.

4.2.3 Discussion

The experiments on this sample of EDF data show there is no definite answer to which pre-processing approach is optimal. None of the mentioned text cleaning strategies provides a clear advantage when using the text for topic modelling, which will be used as one of the main methods we discuss for later experiments in Chapter 5. Thus, it should be noted that while text pre-processing can offer some optimization for NLP tasks, it is also not guaranteed to provide a distinct boost in performance and should be considered as an afterthought when conducting experiments to formulate a research method. As shown in the experiment results, no further pre-processing other than the mandatory ones (boilerplate and financial vocabulary removal) can deliver text quality that LDA can perform respectably on. Ultimately, we have chosen the removal of all special characters, numbers, and stopwords as our pre-processing strategy for the EDF data for subsequent experiments, based on the coherence score results showing that the approach yielded the most discernible peaks compared to the rest.

4.3 Conclusions

In this chapter, we answered the question of the extent to which noise affects tasks and whether pre-processing is a mandatory step. Firstly, in the case of the DAnIEL system and dataset, it is discovered that the epidemical event extraction system is prone to noise, but, at the same time, the impact on the DANIEL system is not considerable, which makes it a robust solution for health surveillance applications. Thus, given an unfavourable situation with less desirable quality on the data, the system can still perform. In the experiment with different types of pre-processing, we conclude that it depends on the nature of the data to decide which type of text is unnecessary for future tasks. Moreover, results from using pre-processed text might not outright be better than the original text. Additionally, this chapter further supports the suggestion that not every data should be fed into a single pipeline of execution.

The following chapter will discuss different ETD research that uses the Event-Driven Feeds and the TRENDnert dataset. Its main focus is word representations and whether their temporal context evolution can be used to detect emerging trends in a timely manner in different domains.

CHAPTER 5

A Study About Microsoft at the Beginning of COVID-19 Pandemic in News Articles

Digital news, through many means of diffusion (online publishing platforms, social media, blogs, etc.) are considerably influential, as they not only shape and form public opinion but can also be a factor in the decision-making process of many industries that use technology to improve activities and performance. Therefore, discovering hidden themes and trends residing in news data is essential to improve analyzing, and managing development directions for many companies. The importance of identifying new trends before they emerge is further emphasized with the changing world surrounding the health crisis triggered by the COVID-19 pandemic.

While we did not have any annotation or ground truth available for the EDF data, we still wanted to evaluate the potential of the dataset for the task of ETD. For this reason, we would need an approach that allows us to verify the result easily. Hence, we proposed an adaptation of existing work in emerging trends detection in scientific papers, called Leap2Trend (Dridi et al., 2019). This approach makes use of different Word2Vec (Mikolov et al., 2010) representations of keywords in different periods to identify pairs of keywords growing closer in terms of similarity,

which signify trend movement. However, instead of a global, static Word2Vec representation, we modified the approach with the use of recent contextual embeddings to adapt the system to perform on a news corpus by taking into consideration the temporality of trending topics. Moreover, because Leap2Trend only considered the writing style of scientific papers, we adjusted the keyword extraction phase. In addition, the approach utilizes Google Trend as a golden standard for evaluation which is straightforward to reproduce, thus making it easy to evaluate the performance of the modified system.

Therefore, in this chapter, the main contributions of our study are: (1) We combine term TF-IDF and LDA for a more precise generation of keywords (bi-grams). (2) We utilize the latest contextual embeddings to represent the real temporality and variation of the semantics during different periods and apply it to the Leap2Trend system to identify pairs of keywords that have the potential to become emerging trends.

Results of this chapter have been published in a workshop paper entitled: "*Contextualizing Emerging Trends in Financial News Articles* (Nguyen et al., 2022a)".

5.1 Related Work

Trend Detection in Formal Datasets The general direction for trend detection in formal text data is to use statistical methods (Daud et al., 2021; Hughes et al., 2020), topic modeling (Behpour et al., 2021a; Bolelli et al., 2009b), and clustering (Linger and Hajaiej, 2020; Liu et al., 2020) Using financial business patents, the research by Lee and Sohn (2017) aimed to identify emerging technology trends by applying latent Dirichlet allocation (LDA) with an exponentially weighted moving average of LDA probability, which affects whether a topic is "hot" or "cold". A refined version of TF-IDF, proposed by Zhu et al. (2019), aims to discover "hot" topics according to "hot" terms based on time distribution information, user attention, and K-means clustering. Unlike previous work that tackled trend detection in official documents and newspapers, others focused on proposing new approaches using research and scientific papers, documents that generally contain citations, and bibliographies that could be considered as additional features (Griol-Barres et al., 2020; He et al., 2009; Nie and Sun, 2017; Xu et al., 2019).

However, exploiting bibliographies can have disadvantages in timeliness and content analysis, as discussed by Dridi et al. (2019). The authors further proposed an approach, called Leap2Trend using *temporal* word embedding that was generated by being trained on the data in an initial period of time and then fine-tuned in the upcoming time frames. This approach also tracked the similarities between pairs of keywords over time, which yielded results suggesting the robustness and timeliness characteristics of the Leap2Trend.

Detecting Trends in COVID-19 Pandemic Recent works have also been conducted within the period of the COVID-19 pandemic to study emerging trends within certain communities and gauge the impact of the outbreak through social media text (Kassab et al., 2020). Santis et al. (2020) employed term-frequency analysis, calculating nutrition and energy metrics, while also using social features in order to extract hot terms and build a topic graph through co-occurrence analysis using Twitter data in Italy. Another research targeted peer-review papers regarding the COVID-19 virus and apply word embeddings and machine learning models to track novel insight surrounding the spreading of the virus (Pal et al., 2021).

5.2 Methodology

G.P. provides financial software tools and enterprise applications such as analytics and equity trading platform, data services, and news to financial companies and organizations. The EDF data is massive and contains more than ten years of news collection that Bloomberg published in multiple languages from 2010 up until the present. Not only does the data include the text, but it also incorporates metadata that holds basic information, such as date of publication, headline, and written language. Moreover, the metadata also encloses knowledge that was derived by their own exclusive algorithms, on which general topics and companies relate to each news article.

5.2.1 General Architecture

Our approach extends the Leap2Trend method proposed by Dridi et al. (2019). The authors proposed an interesting approach to the problem of emerging theme detection. Keywords representations were generated in different periods of time using Word2Vec (Mikolov et al., 2013b).

The embeddings are static and only change when a new set of documents from a new time period is used for generating new embeddings. Afterwards, they assessed the similarity evolution of keyword pairs over time to depict which keywords are trending, thus forming topics based on the closeness between keywords. Furthermore, Leap2Trend also tackled the matter of lacking a gold standard to evaluate the result of trend detection by using Google Trends[1]. Google Trends collects search data of keywords and present them as interest rate over time, starting from 2004 to the present, which can be used to project emerging trends prediction results to gauge the performance of the system.

Nonetheless, within the context of our research, Leap2Trend has some disadvantages that needed to be addressed. First and foremost, Leap2Trend used a straightforward approach to extract the main keywords by inspecting titles of scientific papers for the most frequent bi-grams. The solution is justified by the fact that titles from scientific publications are written with the purpose of being self-explanatory and conveying the methods/problems clearly. The writing style leads to titles often containing a substantial amount of keywords. News articles, on the other hand, have condensed headlines that will only be expanded further in the main content of the documents, where most keywords reside. Thus, using raw frequency to extract keywords is inefficient due to noisy text overshadowing important phrases. Secondly, unlike in the scientific corpus that Leap2Trend used, where the context (which is mostly about the computer science field) is rather consistent, news. However, it can contain numerous subjects ranging from technology, finance, and economy to media. Having global static word embeddings, that have a diversified number of abstract topics, might not yield the best results.

Figure 5.3 describes our approach to the research problem, by modifying Leap2Trend based on the aforementioned disadvantages. We extract keywords using TF-IDF scores and LDA topic generation results while using contextual embedding to replace static word embedding. In order to transform a set of incidents into intervals trend detection analysis, we propose a method with six steps. First, we pre-process our dataset to remove unwanted text in order to focus on the main content of the news article. Moreover, the dataset is divided into monthly sub-corpora. The next step is to identify potential keywords from the corpus. We calculate the TF-IDF value

[1] https://trends.google.com/trends/?geo=US

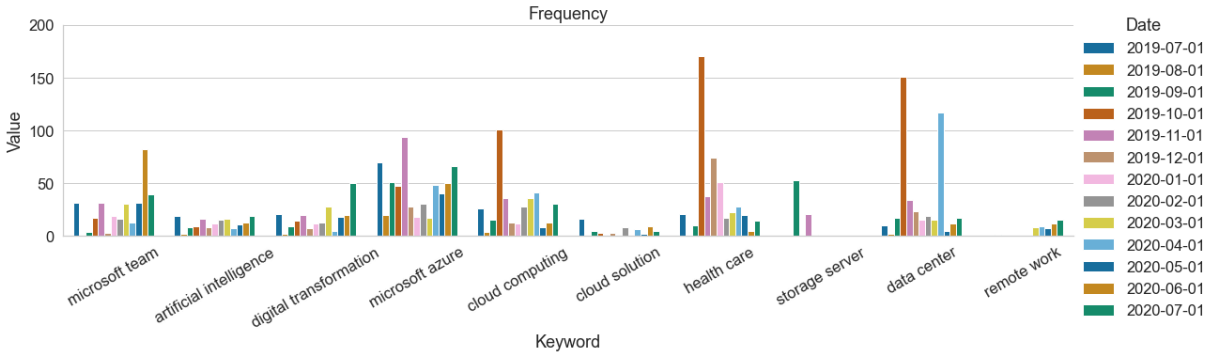


Figure 5.1: Selected potential keywords' frequency.

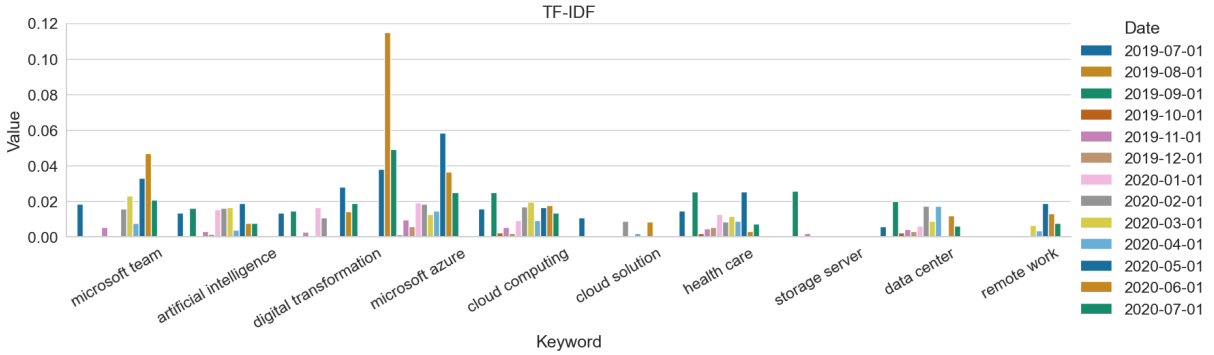


Figure 5.2: Selected potential keywords' TF-IDF value.

as well as apply LDA to generate a list of bi-grams to select from. The detail of this phase is explained further in Section 5.2.2. Afterwards, a list of N keywords is selected from the set of bi-grams generated in the previous step. We hand-picked the keywords ourselves based on our knowledge of the company and the time span of the study. A contextualized representation of the chosen keywords in each monthly sub-corpus was used, which we present in Section 5.2.3. The next step, detailed in Section 5.2.4, is to rank the pair of keywords based on their representation similarity. Lastly, we assess the contextual trend evolution in Section 5.2.5 by analyzing the change in ranking in each pair of keywords over each time span.

5.2.2 Keyword Identification

The research on *keyword identification* discovery originates from the topic detection and tracking (TDT) technology that was first studied by scholars in 1996 and its goal was making new

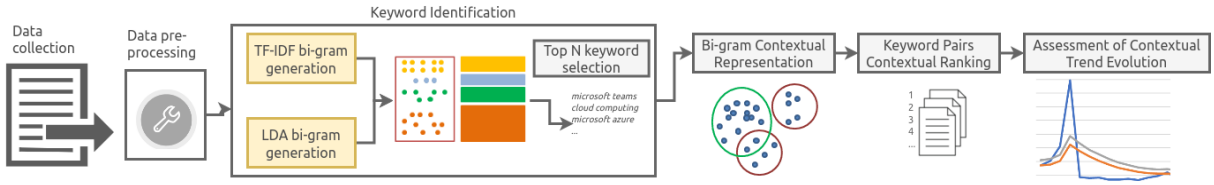


Figure 5.3: Summary of the proposed methodology.

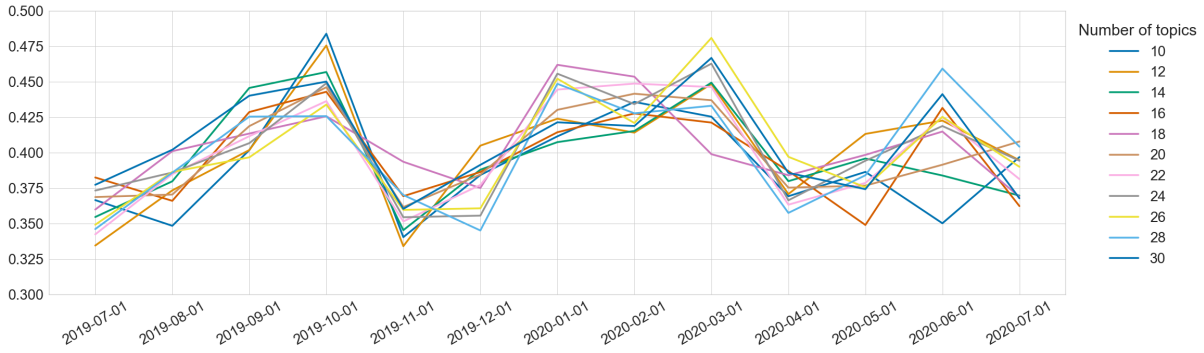


Figure 5.4: LDA coherence score based on the number of topics chosen across every month.

detection and tracking within streams of broadcast news stories (Zhu et al., 2019). Emerging trends are usually signified by terms and phrases that are later considered as defining *keywords* for such themes. Hence, correctly identifying potential keywords will lead in the right direction in discovering promising dormant trends. While keywords can be n-grams, in the scope of this research, only bi-grams were considered due to the fact that compared to uni-grams, they are less ambiguous, while appearing more frequently than other n-grams. Next, we present our methods of extracting potent keywords, using TF-IDF values to generate the list of highly important bi-grams, and utilizing LDA for getting bi-grams that can represent topics.

TF-IDF Bi-gram Generation TF-IDF captures how important a word is in the corpus by considering its frequency and penalizing it for appearing in too many entries in the corpus. Hence, words that are too common have considerably lower TF-IDF values than those that are less frequent. We exploited this method to extract important bi-grams from the corpus. Per month, TF-IDF values are generated for every bi-gram in the news collection and, afterwards, we evaluated and produced lists of bi-grams that have either the highest average TF-IDF value in the collection or the highest single TF-IDF value across all documents. A high average TF-

IDF value may indicate that a keyword associates closely with the company throughout the month, while a single high TF-IDF value signifies a sudden change in the context surrounding the company.

LDA Bi-gram Generation LDA derives from textual data the probabilities of words belonging to a predetermined number of topics. As such, the method excels at providing easily interpretable insights into what consists of a text corpus. Taking advantage of this fact, we used the results of applying LDA on the collection of texts in each month of our dataset to contextualize bi-grams by topics, which we obtained in the previous step. This is done by searching for topics that contain bi-grams in their list of words with the highest probability that represent topics. To find the optimal number of topics, we built numerous LDA models with different values of the number of topics and measured their topic coherence score (Röder et al., 2015). We chose the topic number that gives the highest coherence value. Coherence is a measure to evaluate to which degree the induced topics of an LDA model are correlated to one another, thus choosing the optimal number of topics that marks the end of the rapid growth of topic coherence usually offers meaningful and interpretable topics. While the highest coherence score can be an indicator of optimal value for LDA topic number, we also observed in Figure 5.4 that a larger number of topics could result in generating more duplicated words across multiple topics. Thus, the range of topic numbers was limited with the purpose of distinguishing abstract topics from each other, which in turn yielded higher quality in returned potential keywords.

5.2.3 Bi-gram Contextual Representation Generation

Unlike static word embeddings that capture the global representations of words in a vocabulary, contextual embeddings aim at representing each word or sub-word in the corpus depending on the words surrounding it. Therefore, each appearance of the word will have a unique vector assigned to it, which differentiates the same token but appears in context. For example, the token “teams” in “Microsoft Teams” and “team management” should be represented by distinct vectors, as their context and usage are entirely different. Thus, contextual embedding is generally better than static word embedding, especially in a scenario where the corpus consists of news articles and not documents containing specialized knowledge.

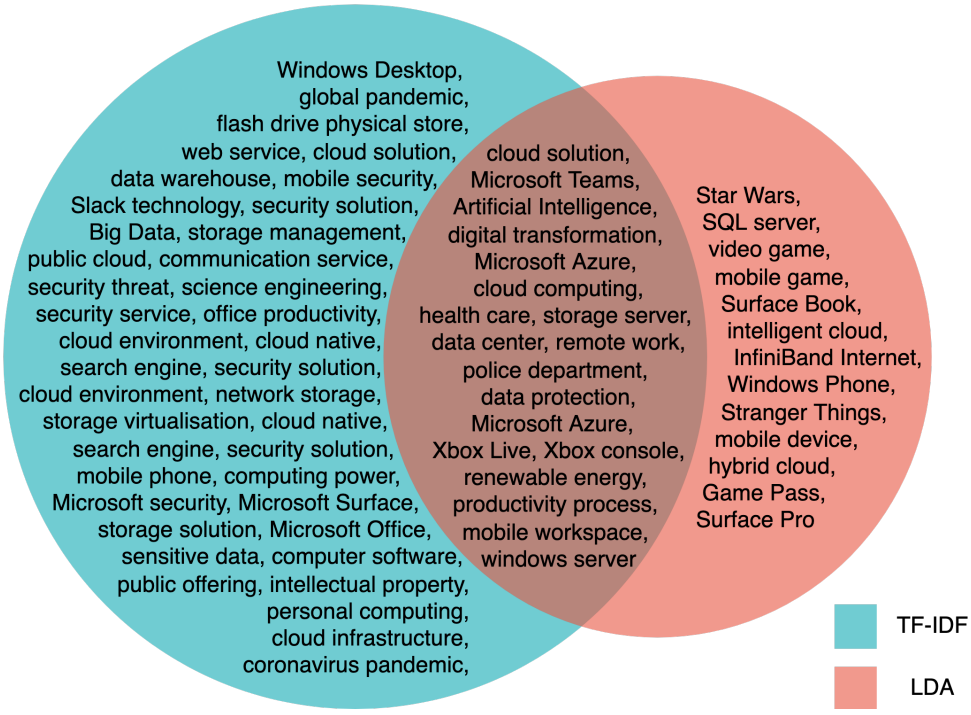


Figure 5.5: Keywords extracted by TF-IDF and LDA.

Because contextual embeddings usually derive vectors for one single token or a character n-gram separately, we employ the following strategy to generate bi-gram embedding to suit our need: for each occurrence of two tokens belonging to a bi-gram appears next to each other and in the right order, we generate FinBERT[2] embeddings for both tokens, and the final vector of the bi-gram is calculated by averaging them.

5.2.4 Keyword Pairs Contextual Ranking

With the contextual embeddings for the set of chosen keywords, we proceed to compute the similarity between each pair of keywords and rank them based on the value calculated. The idea is that when a number of terms appear frequently together, they usually share the same set of surrounding vocabulary, thus having similar context. For this, we computed the cosine similarity. Regarding how to establish the ranking, we first utilized the algorithm employed by Leap2Trend (Dridi et al., 2019), sorting the similarity of pairs of keywords in descending order, where the higher the similarity is, the lower the rank the pair of keywords has.

[2] https://huggingface.co/ProsusAI/finbert

5.2.5 Assessment of Contextual Trend Evolution

Following the ranking calculation for each month, we attempted to assess the contextual evolution of each pair of keywords to identify potential emerging trends relating to the selected keywords. To achieve this, we analyze how much the rank increase/decrease between each month and set a threshold to decide whether changes in ranking signify emerging keywords that can lead to emerging trends. If the differences in ranking between the current month and the previous month of a pair of keywords are greater than the chosen threshold, we identify that this set of keywords will become the emerging terms. On the other hand, if the shift in ranking does not meet the threshold, the pair of keywords is regarded as having no potential in its emergence, either falling off or at standstill in terms of growth for being the next trend.

5.3 Experimental Setup

5.3.1 Gold Standard Creation

To our knowledge at the current time of writing, there is no annotated dataset that is publicly available to experiment with our method on. To produce a gold standard, The process involves examining Google Trends data, a platform for tracking terms/phrases popularity based on Google's search history, of the chosen keywords to identify where their emergence is. This was done by calculating the regression of interest rate evolution from a selected timestamp to N months forward, with a positive value indicating an increase in attention toward the keywords, thus signifying the possibility of the keywords belonging to emerging topics. Formula 5.1 describes the regression of interest rate evolution in the next N months, denoted as m_{hits}:

$$m_{hits} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{5.1}$$

where x_i and y_i represent the month number and the interest rate of that month, respectively. $\bar{x}$ corresponds to the mean of the month number, while $\bar{y}$ is the mean of interest rate.

With the modality Google Trends treats a string in a search query, requesting data on just the literal phrase, we experienced difficulties in extracting interest rate data for combining a pair

of keywords. Thus, we devised a solution by looking for data on each keyword separately and averaging the two results to get a final interest rate for the pair of keywords. The hypothesis behind this is the assumption that both keywords are part of the same theme/topic, thus, their evolution should have a similar tendency to rise/fall, albeit having different magnitude, making the combined signal stays relatively close to the two originals in terms of signal progression. Vice versa, phrases that do not fall into the same category cannot produce a good signal, which automatically makes them irrelevant to each other.

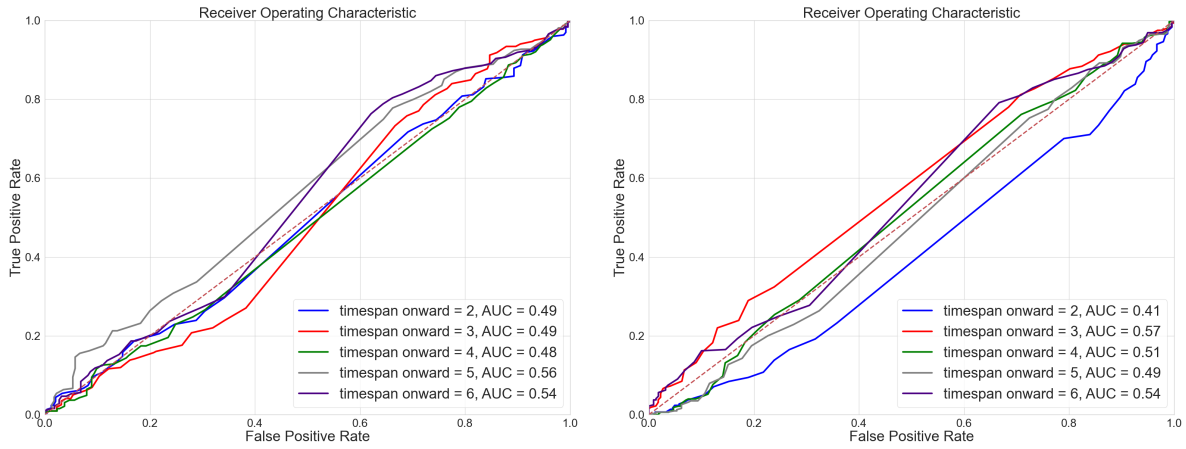
After obtaining the gold standard, we proceed to treat the task as a classification problem, where the system will classify whether changes in ranking context can lead to the same type of movement in the gold standard in the next N months. Accordingly, the main evaluation metrics are Precision, Recall, and F-score. Not only do we consider the macro metrics, but we also take into account the aforementioned metrics on the true class detection specifically, since our focus is leaning toward correctly identifying emerging trends, which is signified by the true class. Additionally, we report the receiver operating characteristic (ROC) curves according to the selected time span onward with the purpose of assessing the detection capability on how many months forward can our system detect trends emergent effectively.

5.3.2 Keyword Generation and Selection

After the keyword generation step (bi-gram generation with TF-IDF and LDA), we noticed that while extracting bi-grams using raw TF-IDF value yielded many potential keywords (68 keywords), the amount of bi-gram that also existed in the LDA results was significantly lower (36 keywords). The intersection of the two lists resulted in a set of 22 keywords. The abundant amount of terms generated by TF-IDF and the high number of intersected keywords between the two methods suggests that results from TF-IDF are less specific than that of LDA. From our observation, the semantics of keywords in the intersect region cover not only the general topics and trends such as health care due to the COVID-19 pandemic but also the specific development direction of Microsoft. Figure 5.5 details further on the list of extracted bi-grams and the total number of bi-grams yielded by each method.

5.4 Results and Discussion

We compare the performance of two systems: the original Leap2Trend which used monthly static Word2Vec embeddings and our Contextual Leap2Trend. We set the threshold=0 for both systems to imply that any positive change in context can signify potential emerging keywords. Figure 5.6b demonstrate the predictive capability of what is the optimal number of months forward the Contextual Leap2Trend respectively can perform the task efficiently. The two systems outperformed one another in different timespan onward scenarios, with the original Leap2Trend having better Area Under the Curve (AUC) when assessing keywords trendiness potential within the span of 5 and 6 months(0.56 and 0.54 in AUC respectively). However, the contextual Leap2Trend system has an AUC of 0.57 when predicting a 3-month forward which not only exceeds the original system in the same category (0.49) but also is the best result overall. This observation suggests that by using contextual embeddings, our system surpasses the original Leap2Trend in terms of timeliness property which is crucial for the task of detecting emerging trends. In addition, the Contextual Leap2Trend matches the original system in AUC (0.54) in the scenario of timespan forward=6.



(a) Original Leap2Trend (b) Contextual Leap2Trend

Figure 5.6: Compared ROC curves based on timespan adjustment (threshold = 0).

While the proposed system is more efficient at predicting trends three months in advance, with the current dataset, the task is considerably challenging. In order to gauge the performance of our system, we compared the results of a system with a threshold of 0 to a zero rule baseline

where every possible bi-gram was considered as trendy (True class), thus True class metrics were not taken into account for this experiment. Table 5.1 shows that our system outperformed the zero rule baseline in all three metrics, but is only marginally better in terms of recall and F-1 value. This observation further supports the difficulty of the task present using the current data.

Another observation is when comparing the change in ranking to Google Trends data, we notice that delays sometimes exists, usually of one month due to data split in the pre-processing phase, in how soon the ranking change with respect to Google Trends, meaning the timeliness aspect of detecting emerging trends might be affected. One possible explanation is that Bloomberg News could be behind in terms of timing compare to public response as Bloomberg mostly covers big events that were already happened, and does not cover innovations processes that can lead to such events.

In the following section, we discuss several example pairs of keywords that signified ongoing trends and emerging ones, mainly about what was happening surrounding the keywords while comparing the graph between Google Trends and the contextual ranking evolution of the Contextual Leap2Trend system.

Table 5.1: Proposed approach vs. zero rule baseline, timespan onward =3.

Method	**Thresh**	**Precision**	**Recall**	**Macro F-score**
Zero-rule baseline		0.3000	0.5000	0.3700
Contextual Leap2Trend	0	**0.5600**	**0.5476**	**0.5388**
Original Leap2Trend	0	0.5384	0.5330	0.5276

5.4.1 A. I. - Digital Transformation

Compared to Google Trends data, for the *artificial intelligence (A.I.) - digital transformation* pair of keywords, our proposed contextual ranking reflects their trends accordingly, as shown in Figure 5.7. Digital transformation and artificial intelligence have been an ongoing conversation in technology and business in recent years as the movement seeks an effort to incorporate A.I. into digitizing business processes, customer experiences, etc. This is illustrated through the European Union's strategy to apply A.I. to digital transformation[3]. As for Microsoft, the com-

[3] https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_264

pany supports this trend with multiple projects, one of them being a major collaboration was mentioned in the EDF data[4].

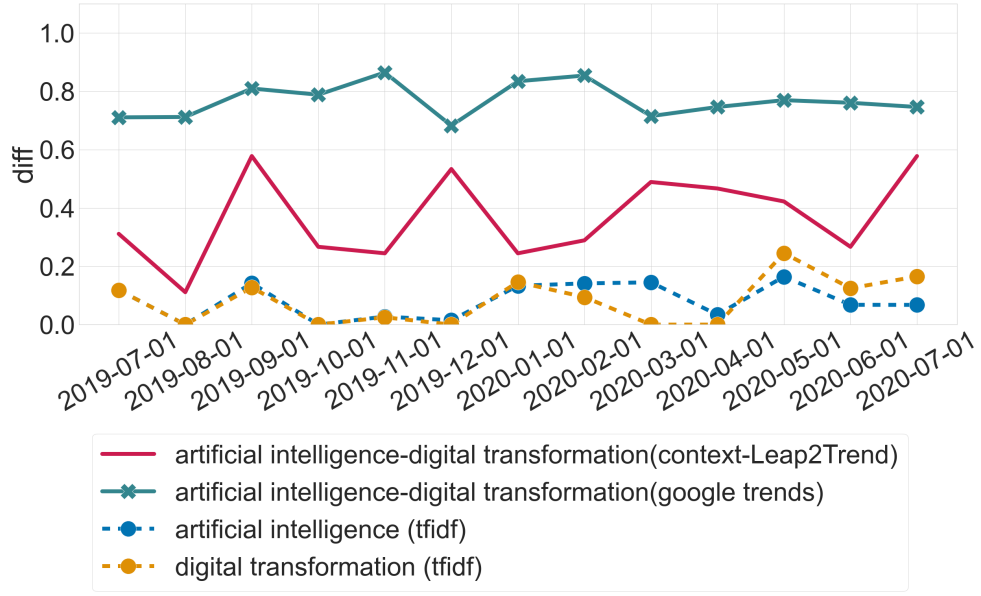


Figure 5.7: Trends for the chosen pair of keywords: A.I. - Digital Transformation

5.4.2 Digital Transformation - Cloud Computing

Within the context of the corpus, the *digital transformation - cloud computing* pair of keywords describes how Microsoft's cloud computing service contributes to their involvement in advancing Digital Transformation with their business partners by providing Azure, Microsoft's leading cloud platform, to enhance the digital capability of their business partner[5]. According to Figure 5.8, the contextual ranking changes of the pair, in general, are in-line with Google Trends data, albeit displayed some level of differences.

[4] https://www.bloomberg.com/press-releases/2020-03-26/c3-ai-microsoft-and-leading-universities-launch-c3-ai-digital-transformation-institute

[5] https://www.bloomberg.com/press-releases/2020-04-07/blackrock-and-microsoft-form-strategic-partnership-to-host-aladdin-on-azure-as-blackrock-readies-aladdin-for-next-chapter-of

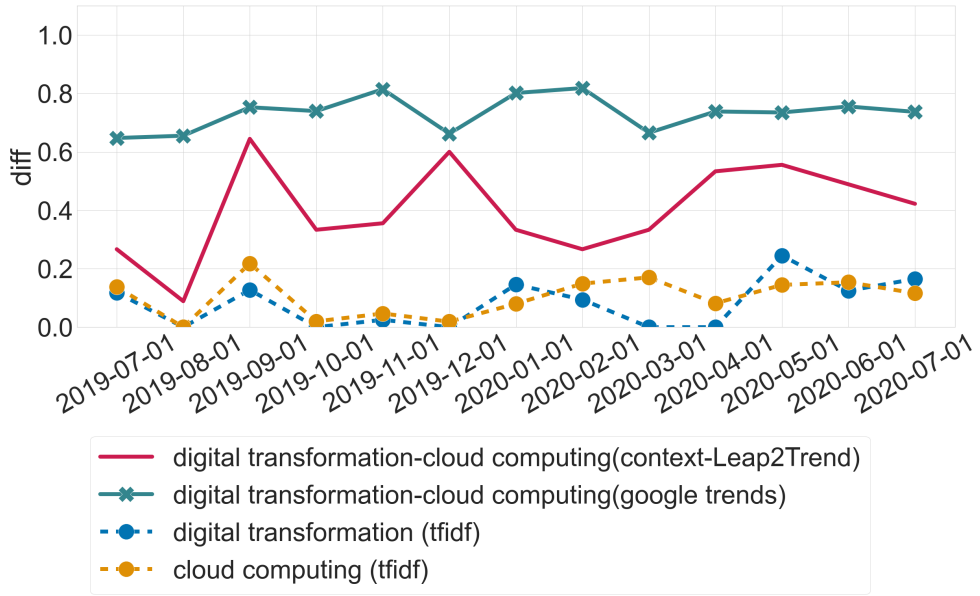


Figure 5.8: Trends for the chosen pair of keywords: Digital Transformation - Cloud Computing

One thing to note is that it is the consensus, mentioned throughout the COVID-19 period in our corpus, regarding *Digital Transformation* that because of the pandemic pushing society to stay distant and work remotely, the *Digital Transformation* process will need to be developed faster to adapt to the current situation. In Figures 5.8 and 5.7, it can be seen that this opinion was reflected through an increase in rankings starting in March 2020. This period is also when the interest rate on Google Trends on this matter started to increase again after a dip.

5.4.3 Microsoft Teams - Remote Work

While the magnitude displayed in Figure 5.9 was definitely lower than Google Trends's signal, Leap2Trend's signal visually still showed signs that the trends of the *Microsoft Teams - remote work* pair of keywords have potential. Remote work, while being lesser known in 2019, has been a staple since the COVID-19 pandemic began. Zoom, a platform for online meetings, was booming at the start of the pandemic, yet fell in popularity due to security reasons. The slump of Zoom paved the way for the rise of Microsoft Teams as the reliable platform for workplace communication[6]. In our EDF dataset, the development of Microsoft Teams with new and better

[6] https://www.computerweekly.com/news/252485100/Microsoft-Teams-usage-growth-surpasses-Zoom

features also aid in its popularity[7].

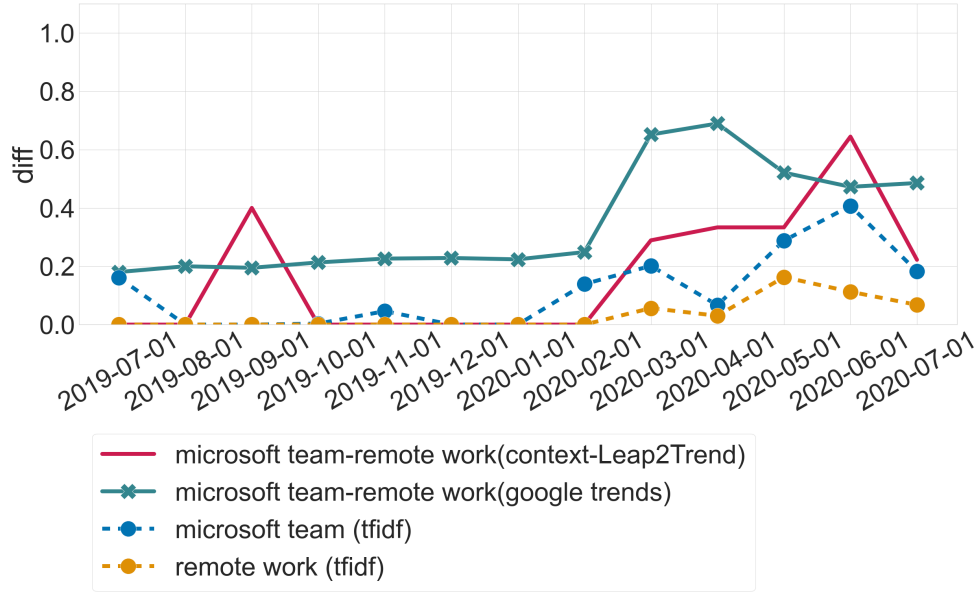


Figure 5.9: Trends for the chosen pair of keywords: Microsoft Teams - Remote Work

5.5 Conclusions

This research addressed the drawback of existing methods in keyword extraction and bi-gram representation due to the differences in writing style between scientific papers and news articles. We, instead, introduced a combination of TF-IDF and LDA for generating potential keywords and utilized contextual embeddings for the change in temporality. Our proposed system showed considerable improvements compared to the original method in some scenarios in length of prediction. Moreover, we also presented several examples of emerging trends found in our data and the result also suggested that the approach has a good timeliness characteristic In future work, we plan to introduce a better variety of data, such as news articles covering more companies and sectors, to further experiment and improve our system. Moreover, increasing the time intervals could also uphold the consideration to assess the trend's longevity.

The next chapter presents a continuation of the research surrounding the context evolution of a text span. However, instead of assessing pairs of keywords, we focus on separated keywords and

[7] https://www.bloomberg.com/news/articles/2020-03-19/microsoft-teams-boosts-work-at-home-effort

evaluate their changes in context by comparing current points to a continuous past time period rather than just two consecutive points in the timeline. Moreover, we will be moving from the news articles EDF dataset to the dataset of computer science publications TRENDnert.

CHAPTER 6

Detecting Up and Emerging Trends with Long-term Impact in Scientific Articles

This chapter focuses on exploring the dynamics between how the context of keywords evolves in multiple points of reference and the emergence of related trends in the scope of scientific publications. While previous works that dealt with scientific publications for ETD have been incorporating citations and references as the main feature to track trend evolution (Griol-Barres et al., 2020; He et al., 2009; Nie and Sun, 2017; Xu et al., 2019), it would take a long period of time for these two metrics to fully establish, thus trend analysis and detection can be lacking in terms of timeliness (He and Chen, 2018). Moreover, research in ETD revolving around text and context similarity did not include temporal factor (Akrouchi et al., 2021; Daud et al., 2021) or introduced too few points of reference in time for comparing with the current time span (Dridi et al., 2019; Linger and Hajaiej, 2020).

Therefore, our research focuses on exploring the dynamics between how the context of keywords evolves in multiple points of reference and the emergence of related trends in the scope of scientific publications. The main contribution of the paper is as follows: (1) we introduce a

new concept called the *Term Context Evolution*, in which we extract the term representations of each time period from a pre-trained Transformer model and create a series of representation similarities between the current time span and a pre-defined number of past periods. The goal of forming this series of similarity values is to track how different a term is from its past self, thus drawing a conclusion about whether the term in the current self has the potential to emerge into a new trend. (2) Using the context evolution of key terms, we attempt to detect their up-trend movement, where they are considered growing in popularity. Moreover, we are interested in identifying periods that are considered the point of emergence for keywords using their change in the surrounding context.

The work of this chapter has generated the following publication: "*Utilizing Keywords Evolution in Context for Emerging Trend Detection in Scientific Publications*" (Nguyen et al., 2022b).

6.1 Related Work

Research in ETD apply various types of methods such as statistical methods (Daud et al., 2021; Hughes et al., 2020; Zhu et al., 2019), graph-based methods (Dang et al., 2016; Pal et al., 2021), topic modeling-based methods (Akrouchi et al., 2021; Behpour et al., 2021a; Bolelli et al., 2009a), clustering methods (Li et al., 2020a; Qiu et al., 2021), or some combinations of the above (Lee and Sohn, 2017; Liu et al., 2020; Santis et al., 2020). Selecting an approach also relies on the availability of external, content-independent features that measure the diffusion rate of documents. For example, social media counts the number of likes, shares, and comments while a news article has a specific number of views/hits to the site.

Scientific publications use citation-based metrics to gauge the popularity and impact of the research by considering how many works reference the paper after release. Combining with text mining methods, bibliography analysis has been utilized in various approaches, such as estimating how influential and useful a paper is by Nie and Sun (2017), creating a citation network to track novelty in research topics (Xu et al., 2019), using h-index to indicate the degree of transmission of a paper (Griol-Barres et al., 2020) or complementing existing methods as extra arguments (He et al., 2009), to track the evolution of prominent research trends. However, be-

cause of the time needed for the citation list of a paper to be fully developed to a certain degree, the lack of timeliness is a disadvantage when using citation-based metric (Dridi et al., 2019).

Text analysis, unlike citation analysis, has less time lag since the text is available immediately and does not need time for a list of citations and references to be fully developed. Hence, directly exploring emerging trends through the content of documents is the better approach considering timeliness is often a major factor in ETD (He and Chen, 2018). The approach of using a similarity metric to track the convergence of smaller topics into one emerging topic has been studied in previous research, including closeness assessment of clusters through time and linking clusters with the highest similarity together to form a topic (Li et al., 2020a; Linger and Hajaiej, 2020), exploring words to support weak signals with similarity of Word2Vec representation (Akrouchi et al., 2021), and creating graphs of relating entities to detect the emergence of keywords (Pal et al., 2021).

Yet, for scientific corpora specifically, exploiting similarities to detect the emergence of trends is still an underdeveloped approach. Recently, attempts have been made to adopt similarity analysis of temporal embeddings, generally, Word2Vec embeddings trained on documents from different time periods, to assess the novelty of detected keywords in published papers (He and Chen, 2018), or, in the example of Leap2Trend discussed in Chapter 5 (Dridi et al., 2019), to evaluate potential pairs of keywords that might grow into a trend. While the results of previous ETD research utilizing similarity showed some degrees of success, the aforementioned methods only considered the similarity of keywords between two consecutive time intervals, which can be too short to fully gauge the evolution of context surrounding keywords, thus results can be inconsistent. In addition, as also mentioned in previous chapters, Word2Vec provides static word embeddings that are context-independent, thus, suffering from a major drawback of lacking context in the representation and potentially losing information.

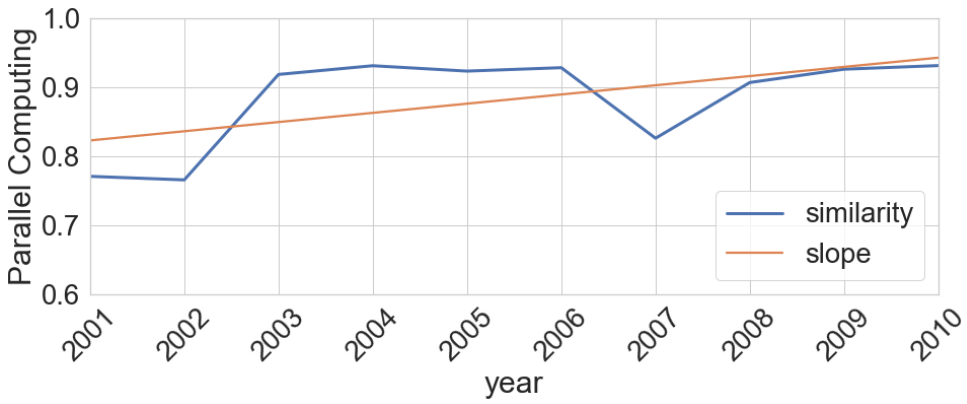


Figure 6.1: The evolution of "Parallel Computing" context by comparing 2011 representation to itself from 2001 to 2010.

6.2 Methodology

6.2.1 Term Context Evolution Approach

In this section, we introduce the *Term Context Evolution* (TCE) approach. As previously mentioned, context change of words/phrases regarding representation similarity between two successive time periods can be insufficient to extrapolate evolution leading to topics emergence. Hence, the main objective of TCE is to have a stronger and more complete analysis of context evolution by contrasting the current representation to that of not only the most recent but also a considerable number of past periods.

Definition Given a term, within a time span of $T = \{T_0, T_1, ..., T_N\}$ where T is a time period measured in months, or years, and α is the length of time span to look into the past, TCE at point-in-time T_i is defined as followed:

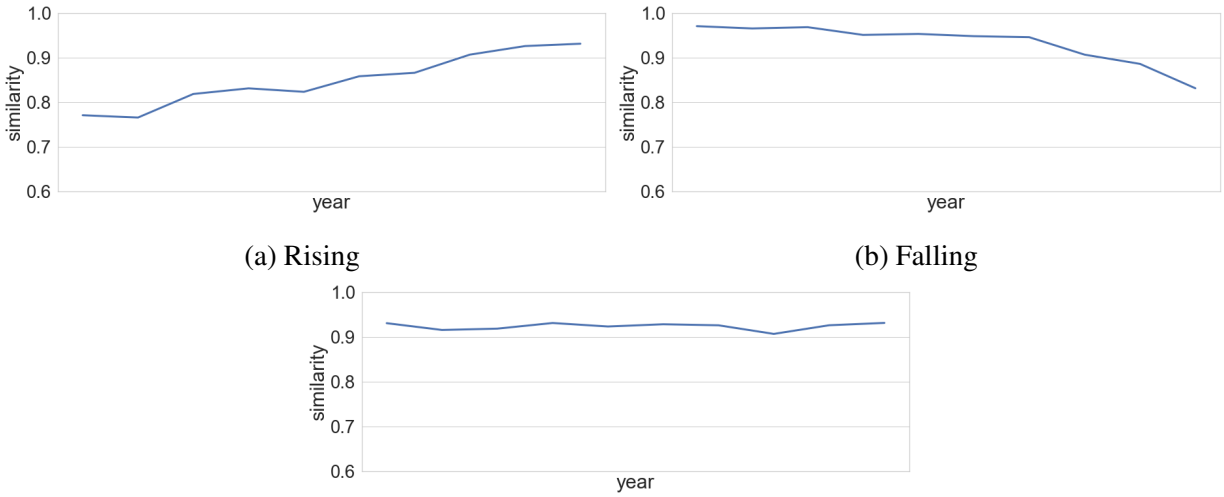
$$C_i = \{c_{i-\alpha}, c_{i-\alpha+1}, ..., c_{i-1}\}$$

where

$$c_{i-k} = cosine\_similarity(w_i, w_{i-k})$$

with $k = [1, ..., \alpha]$, w_i and w_{i-k} are the representations at point-in-time T_i and T_{i-k}, respectively.

The expected output is a series of similarity metrics between the current representation of a term and its past representations. The main objective of using TCE is to express the change in the context of terms and keywords throughout the year, thus deducting possible situations regarding the moving direction of the term or keyword. Figure 6.1 shows an example of TCE on the term "parallel computing" in 2011, with the y-axis being the similarity value of representation between 2011 and each year in the x-axis.



(a) Rising

(b) Falling

(c) Stagnating

Figure 6.2: Example of three main different behaviors of our *Term Context Evolution* approach of a given month.

Figure 6.2 gives an insight into the three main possible situations our TCE approach can give. In Figure 6.2a, the context of a given term is evolving from being different in the past to having quite a similar context compared to the current representation, thus signifying a change in usage which is accompanied by an increase in usage and potentially popularity gain or trend emergence. In contrast, the reverse situation is illustrated in figure 6.2b where the context is more similar to the past rather than recent periods, which implies that the term has had some changes in usage, yet returns to be more similar to its past self. Lastly, Figure 6.2c describes the state of stagnant as the context of a term does not change throughout the course of a given interval.

6.2.2 Slope Calculation for Emerging Trend Prediction

In order to assess the slope direction of the TCE at T_i, Equation 6.1 describes the linear regression formula that is applied to calculate the context evolution slope of the series.

$$slope_i = \frac{\sum_{k=-\alpha}^{-1}(x_{i-k} - \bar{x_i})(c_{i-k} - \bar{c_i})}{\sum_{k=-\alpha}^{-1}(x_{i-k} - \bar{x_i})^2} \tag{6.1}$$

where $\bar{c_i}$ is the mean of the series, C_i, x_{i-k} is the year within the period to look into the past and $\bar{x_i}$ is the mean year of the period.

As mentioned in the previous section, two tasks are of interest: uptrend detection which is defined as the identification of growing movements and distinguishing them from stagnate or declining trends, and point of emergence identification which is the task of finding the periods where a topic has the potential to become a trend, hence the emergence.

The main objective of calculating the slope direction of the TCE is to utilize it as a basis for our prediction given the two aforementioned tasks. If $slope_i$ is within a specific threshold range, the term should be considered as a candidate at said time, otherwise marked as a non-candidate. For the task of uptrend detection, we set the threshold to 0, thus establishing any positive change in similarity would result in a potential trending growth. In contrast, since the point of emergence identification deals not only with trend growth but also the first period where the trend emerges, the threshold is tightened with an upper bound of 0.1. If $slope_i$ crosses the upper-bound value, we observed that terms in that period have already gone past their emergence phase, thus making them not of our interest.

6.3 Experimental Setup

6.3.1 Terms Selection

To select suitable terms for the experiment, we adopt the approach of selecting terms based on their total number of appearances in the title of every publication. As the title of a scientific paper usually tries to encapsulate the content in a few words, keywords that summarize the pa-

Table 6.1: Selected terms and their corresponding topics.

Term	Topic Mapping
neural networks, artificial intelligence	AI: Machine and Deep Learning
face detection, object recognition	Computer Vision, Object Tracking, and Face Recognition
transfer learning	Online Learning
multimedia application	Multimedia
energy consumption	Energy Conversion and Power Electronics
genetic algorithms	Genetic Algorithms
virtual machines	Cloud Computing and Software as a Service
information extraction	Information and Knowledge Management
parallel computing	Parallel Computing
data analysis	Big Data Analytics
ray tracing	Technologies for Computing Hardware: GPU and SpeedUp
fault detection	E-Business and E-Commerce: Patents
quantum computing	Quantum Computing

per well are likely to be included. Thus, for our experiment, we select the 15 most occurred bi-grams to assess their context evolution and retrospectively evaluate their potential emergence as a trend. The chosen bi-grams are as follows: neural network, face detection, transfer learning, artificial intelligence, object recognition, multimedia application, energy consumption, genetic algorithms, virtual machines, information extraction, parallel computing, data analysis, ray tracing, fault detection, and quantum computing.

As previously discussed, the TRENDNert dataset contains only annotated information on the general topics, it is required to map our selected keywords to the corresponding topics for evaluation. Hence, we manually mapped the selected keywords to their most related trending topics. Table 6.1 shows the terms and their respective topics that appear in TRENDNert.

6.3.2 Extracting Contextual Representation of the Terms

The main objective of the research is to evaluate the evolution of context surrounding important keywords, thus we extract the term representations using SciBERT (Beltagy et al., 2019), a BERT-based Transformer model that was pre-trained on scientific papers. The process of obtaining the embeddings of selected terms is as followed. In each period, every abstract is split into sentences, which are then fed into the tokenizer and Transformer pre-trained model to get the representation of each token in the abstract. Using the same tokenizer, the keywords are

transformed into a list of tokens. Afterward, we extract from the tokenized text every sequence of tokens that match those of selected keywords. For each token in the extracted sequence, the last hidden layer in its representation in the model is chosen as the token vector. To achieve the representation of one appearance of a sequence, we take the mean of the vectors of tokens in the sequence. Finally, an average of every representation of a keyword is performed to generate the final vector of a period.

6.3.3 Ground Truth and Evaluation

To create a ground truth for our experiments, we relied on topic popularity, which is the number of papers regarding each topic published throughout the year. Using linear regression as presented in Formula 6.1, we calculate the growth/stagnation of each topic every three years forward with respect to the current point in time. The number of years was chosen since we want to detect the long-term trends and estimate that three years is a good amount of time for trends to develop. The main reason we adopted this metric is that it is straightforward and simple to gauge the movement surrounding topics using a number of published research, where the more research conducted relating to one subject, the more popular it can be, hence creating a trend of research.

The linear regression formula to calculate the growth/stagnation rate of topics in three years onward at a given point of time is as followed:

$$growth\_rate = \frac{\sum_{i=0}^{3}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^{3}(x_i - \bar{x})^2} \tag{6.2}$$

where x_i and y_i represent the year and the number of published papers on a given topic for that year, respectively. $\bar{x}$ corresponds to the mean of the year number, while $\bar{y}$ is the mean of publishing frequency.

As previously mentioned, we are interested in discovering the relation between context change and trend, which is reflected in two tasks:

1. *Uptrend detection*: In order to evaluate the task, we created the ground truth by setting the lower-bound threshold to 2.5, as a lower growth rate value usually indicates a minor

change in popularity, thus not significant enough to conclude the growth of a topic.

2. *Point of emergence identification*: In contrast to the first task, the threshold was tightened, not only placing the lower-bound threshold at 2.5 but also setting the upper-bound to 10 with the intention of limiting which range of growth rate can signify a topic emergence. In our observation, topics of the current point in time with a growth rate that is greater than 10 usually have gone past the emergence phase and have transited into the booming phase. Because of stated reasons, having ground truth outside the selected defeats the purpose of detecting emergent points of topics.

6.4 Results and Discussion

We compare the result of up-trend detection of two systems: our proposed approach and state-of-the-art system referred to as Leap2Trend (Dridi et al., 2019), using the hyperparameters suggested by the authors of the research.

While Leap2Trend also explores the use of temporal similarity between vector representations of keywords, the approach only compares the nearest interval with the current ones to determine what could be in trend. Moreover, as mentioned in the previous section, static embeddings represent text globally without considering the context the text appears in, thus can potentially cause a loss of information. Because of the aforementioned similarities and differences, we decided to choose Leap2Trend as a reference system.

Table 6.2 demonstrates the predictive capability of two approaches in terms of Precision, Recall, and F-score. We also report the macro-average (macro-AVG) for both uptrend and non-uptrend classes. A macro-average computes the metric independently for each class and then takes the average, hence treating all classes equally.

Overall, it can be observed that our proposed approach exceeds Leap2Trend when predicting uptrend class with superior recall whereas having marginally higher precision. Upon closer inspection, the recall value of our method is exceptionally higher than the precision value with uptrend prediction while the reverse occurs with the non-uptrend class.

Table 6.2: Comparing performance on uptrend prediction.

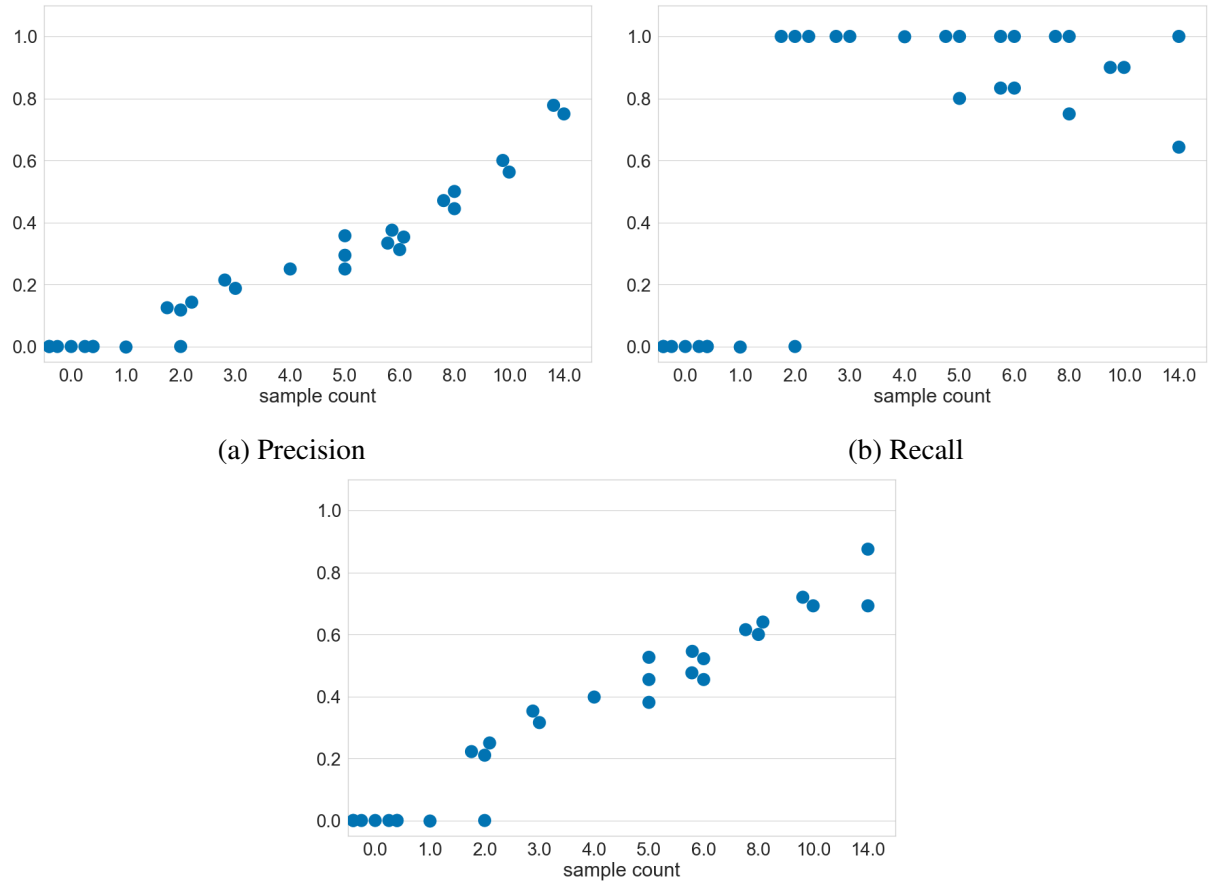
		Precision	Recall	F-score
Non-Uptrend	TCE	**81.11**	19.95	32.02
	Leap2Trend	71.43	56.01	62.79
Uptrend	TCE	**31.81**	**89.10**	47.28
	Leap2Trend	31.49	47.44	47.85
Macro AVG	TCE	**56.64**	**54.52**	39.65
	Leap2Trend	51.46	51.72	50.32

With the task of emerging point identification, the proposed approach performs slightly better on non-emerge class compared to that of non-uptrend class prediction in the previous task. On the other hand, results on the emerging class are considerably lower[1].

Table 6.3: TCE approach performance on emerging point prediction.

	Precision	Recall	F-score
Non-Emerge	85.20	38.84	53.35
Emerge	19.33	68.48	30.14
Macro-AVG	52.26	53.66	41.75

[1]It should be noted that in this experiment, we decided to exclude Leap2Trend due to the task not being in the scope of its main objectives.



(a) Precision

(b) Recall

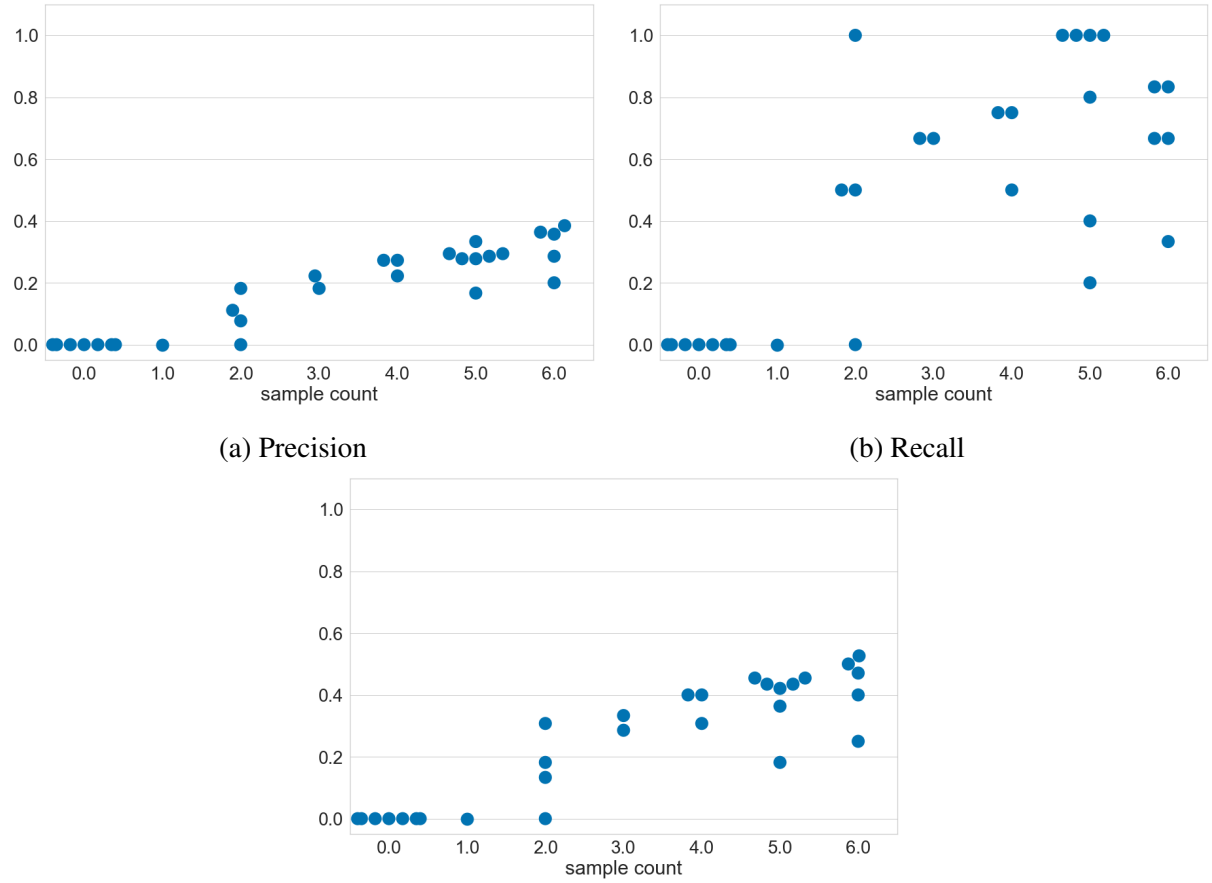
(c) F-score

Figure 6.3: Uptrend Prediction performance distribution per number of support of term-topics

6.4.1 General Analysis

Figures 6.3 and 6.4 describe the relationship between the number of true positives and the performance of the system. Ignoring cases where the true positive value is 0, the general tendency for both tasks is that the results get better as the sample count gets bigger, with high to very high recall, as illustrated in Figure 6.3b and 6.4b. However, in the case of uptrend detection, the pattern is more recognizable while it is more erratic with emerging point identification. An observation in Figures 6.3a and 6.4a to note is that precision in true positive prediction in both tasks correlates positively with the number of true positives, which leads to the same behavior in F-score, shown in Figures 6.3c and 6.4c. All things considered, the high recall results serve better for the task of emerging trend detection since it is still the expert that makes the final

decision on which topics to invest in/trust (similar to a recommendation system). Thus, getting as many true positives as possible while keeping a respectable precision/recall should be what to aim for.



(a) Precision

(b) Recall

(c) F-score

Figure 6.4: Emerging Point Prediction performance distribution per number of support of term-topics

6.4.2 Trend Prediction Cases

From our experimental setup, the results revealed notable insights in regard to the terms presented in Table 6.1.

Neural networks and **artificial intelligence (AI)**: regarding related labels, they started gaining attention in the 90s in the real world (Liu et al., 2018), with impactful events such as AI de-

feating a reigning chess champion in 1997[2]. This uptrend movement for these keywords was also reflected in our result, where throughout the respective timeline in the dataset, the system considers the emergence of these keywords at many different points in the 90s.

Face detection and **object recognition**: the annotated data acknowledge relating labels as downtrends. However, the results showed a different story as both were predicted to emerge in the late 2000s, which happened later in the 2010s with to the advance of computer vision (CV) due to the emergence and popularity of neural networks and AI as stated previously (Insaf et al., 2020; Zou et al., 2019).

Multimedia applications became a downtrend in the annotated dataset with no point of emergence and it shows in the point of emergence prediction results. The uptrend prediction result with this term was stagnant, thus further strengthening the observation.

Virtual machines: we detected its emergence point during the 2000s, which aligned with the release of Amazon Web Service (AWS) earlier that time frame. AWS's development was influential, causing the rise in cloud computing, where virtual machines played a crucial role in its evolution, in the following years until the present[3].

6.5 Conclusions

We tackled the task of emerging trend detection by utilizing context evolution of terms to track periods where significant change usage can indicate potential forthcoming as a trend of said terms. We, thus, introduced the *Term Context Evolution* approach in order to explore how similar a word at a certain period of time is to its past self, by formulating a basis for prediction regarding uptrends and points of emergence.

Using a Transformer-based pre-trained model to contextually represent scientific texts from the TRENDNert dataset, our system was able to outperform existing methods in terms of true class recall in uptrend detection while maintaining a similar level of precision. Moreover, the proposed method also had respectable results in discovering points of trend emergence.

[2] https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence
[3] https://www.dataversity.net/how-the-cloud-has-evolved-over-the-past-10-years/

Finally, we presented several examples of emerging trends found in the data that suggested that our approach has a good timeliness characteristic. In addition, the high recall in our predictions can be useful as suggestions for experts to make decisions in order to adopt topics as investments or new research directions.

In future work, we intend to expand the variety of data to include other fields of study aside from computer science. Finding an annotated dataset that is designed explicitly for the task of emerging trend detection is also a crucial part of our upcoming research. Additionally, more research is needed to fully explore the potential of our proposed TCE approach in regard to hyperparameters selection, and analysis methods.

CHAPTER 7

Hypernym Detection using Sentence Transformers (FinSim 2021)

As established in previous chapters, the datasets that are available to us do not have the proper annotation to fully evaluate our method. Because of this, we decided to participate in an evaluation campaign called FinSim-2, which focused on the task of hypernym detection in the context of financial vocabulary. We believe that this shared task was suitable for our research for the following reasons: Firstly, the task was presented in a financial setting, which caught our attention and could be beneficial to the "Smart Beta" team of LBP-AM. Secondly, the organizers provided a fully annotated dataset with a respectable number of terms and their abstract concepts, which allowed us to evaluate our approach to terms similarity. Lastly, terms were mapped manually to a topic in previous chapters, which raises the problem of automatically detecting a category a term belongs to and vice versa, further generalizing/specifying topics depending on the need to grasp a better view of what is trending or what is causing the trend. Hence, we deemed that the campaign would also help us in connecting the terms we identified as emerging to form grander, broader emerging themes to take into account rather than just separated keywords.

While there exists a few predominant applications of natural language processing (NLP) regarding finance, for example analyzing the sentiment of financial news or reports, many practices remain under-represented. One of these unexplored research topics is hypernym extraction. A hypernym indicates a word/concept that has the highest level of category abstraction and generally has hyponyms, which are terms describing a more specific concept in said category. The task of hypernym extraction generally deals with finding the hypernym-hyponym association that usually belongs to the kind of "is-a" relationship.

To the best of our knowledge, FinSim[1] is the first shared task that tackles hypernym extraction in the financial domain. Rather than using news or annual reports, the organizers targeted prospectus (Mansar et al., 2021), a type of financial document that describes investments offered to the public, which is mandatory to file and submit to the Securities and Exchange Commission. The shared task consists of a list of financial terms extracted from a set of prospectuses that need to be assigned to their corresponding top-level concepts. These hypernyms are known beforehand, hence the task can be considered a multi-class classification.

The experiments in this chapter have been presented in a publication entitled: "*"L3i_LBPAM at the FinSim-2 task: Learning Financial Semantic Similarities with Siamese Transformers*" (Nguyen et al., 2021).

7.1 Related Work

According to a recent survey by Wang et al. (2017), hypernym extraction consists of two general approaches: pattern-based and distributional-based. Pattern-based methods are traditional approaches to this problem that attempt to find the pair of terms that satisfy certain patterns. For example, a method that analyzed the co-occurrences of words to discover hyponym-hypernym couples was proposed (Grefenstette, 2015). With distributional methods, they have been given more attention recently with the advances of word embeddings such as Word2Vec (Mikolov et al., 2013b), and GloVE (Pennington et al., 2014) (context-independent), or BERT (Devlin et al., 2018) (context-dependent).

[1] https://sites.google.com/nlg.csie.ntu.edu.tw/finweb2021/shared-task-finsim-2

Word embeddings can capture different similarities between terms and their top-level concepts. For example, the SentEval[2] toolkit evaluated the semantic similarities between texts (Conneau and Kiela, 2018). Another notable research designed a siamese network using BERT as an encoder in order to measure similarities between sentences (Reimers and Gurevych, 2019).

7.2 Methodology

The architecture we proposed is based on a siamese neural network that contains two pre-trained BERT encoders proposed by Devlin et al. (2018) with the same configuration and the same hyperparameters[3], as presented in Figure 7.2. This type of architecture allows updating the weights of both encoders such that the produced term embeddings are semantically meaningful and can be compared. We use cosine similarity for comparing the representations between a given term and the ontology term.

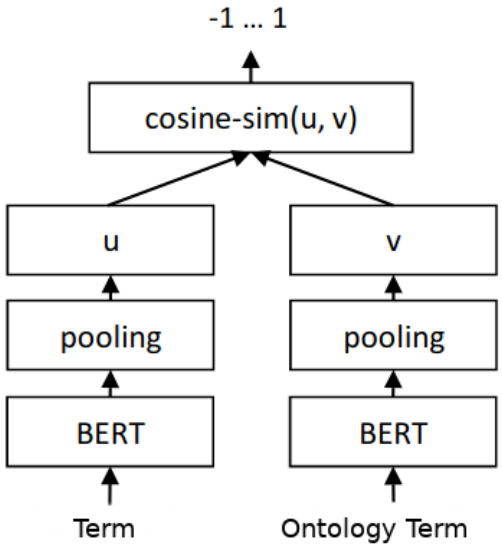


Figure 7.1: Architecture on calculating semantic similarities using a siamese network with BERT encoders (Reimers and Gurevych, 2019).

The model adds a pooling operation to the output of the BERT encoder using the output of the

[2] https://github.com/facebookresearch/SentEval
[3] The method is implemented and available at https://github.com/UKPLab/sentence-transformers

[CLS]-token and we computed the mean of all output vectors, as this strategy proved the best results in our preliminary experiments.

7.3 Experimental Setup

7.3.1 FinSim Data

For training the models, we were provided with 600 terms along with their respective hypernyms/labels, which include ten top-level financial concepts: *Bonds, Forward, Funds, Future, MMIs, Option, Stocks*, and *Swap*.

Upon observation, we noticed several challenges. First, the training dataset is relatively small, with only 600 entries. While it is much larger than last year (100 terms for the training data), it is still not enough to apply neural-based approaches. As for the characteristic of the terms given in the training set, we realized that, occasionally, these terms contain the top-level concepts to which they belong. Most of the time, terms usually do not have any hypernyms as part of them. However, there are terms that contain hypernyms of other classes while belonging to different top-level concepts. In some extreme cases, a term can have both hypernyms that it belongs to and hypernyms that are irrelevant. A number of 119 related hypernyms was observed, while the number of unrelated hypernyms was 53, and 12 were common to both, in a total of 160 hypernyms.

Here are a few examples extracted from the dataset: "Corporate Bonds" contains the "Bonds" hypernym, which is also the category this phrase belongs to. With "Fixed Recovery Swap", while it has the "Swap" concept in it, this term is not a lower concept of said hypernym, but belongs to "Credit Index". Including the "Bond" and "Future" hypernyms, the term "Single Name Bond Future", however, has a top-level ontology term "Future".

7.3.2 Data Pre-processing

In order to decide which are the most performing systems, we split the provided annotated data into 80% for training and 10% for each development and test set. For every top-level ontology

term (10), the input entry is (term, ontology term, distance). 0 means that the term is irrelevant to the ontology term and 1 indicates a close relationship between the term and the hypernym, with a maximum distance of 1.

(1) *Credit Default Swap.*, Swap, 1

(2) *Credit Default Swap.*, Credit Index, 0

(3) *Credit Default Swap.*, Bonds, 0 . . .

7.3.3 Detecting Hypernyms

In this section, we describe in detail the architecture of the baseline and the more sophisticated systems we employed.

Baseline We investigated four commonly used text classification models as baselines: logistic regression, random forest, support vector machine, and decision trees. For all models, we used the default hyperparameters and a TF-IDF (term frequency-inverse document frequency) weighting measure[4]. We also consider as features two different embeddings: the GloVE (Pennington et al., 2014) embeddings and the *FinSim* embeddings provided by the organizers that were trained on a set of financial articles.

Metrics Regarding what metrics to evaluate the results, we followed the two criteria that FinSim organizers employed: *Mean Rank* and *Accuracy*. Accuracy implies the rate of correct prediction from the system compared to the ground truth. On the other hand, the *Mean Rank* expresses how far off the correct label was in the prediction from the first rank. However, to ensure that this metric stays consistent, the organizers imposed a limit where if the correct label is not in the top 3 of the prediction, the rank of that prediction is automatically assigned as four regardless of how low the label is.

Hyperparameters For our baseline models, we used the default parameters. For the pretrained encoders, we experimented with several English ones (`bert-base`, `bert-large`,

[4] https://scikit-learn.org/

`bert-STS`[5] *cased* and *uncased* models). We trained for 150 epochs, with Adam optimizer with weight decay, 2×10^{-5} learning rate, and a mini-batch of dimension 8.

7.4 Results and Discussion

Table 7.1 describes the results of different methods on the test dataset that was split from the given 600 terms. We used different machine learning techniques as baselines: logistic regression, support vector machine, random forest, and decision trees. Unigram TF-IDF along with two pre-trained embeddings, the provided FinSim embeddings, and GloVE[6], were used. Among these baselines, logistic regression and SVM coupled with GloVE yielded the best results, where SVM performed better in the mean rank metric while logistic regression has better accuracy by a small margin.

In the results from Table 7.1 we remark that the chosen baseline models with TF-IDF weighting, GloVe, and FinSim embeddings are competitive with the siamese-based models with pre-trained BERT encoders. We also notice that using FinSim pre-trained embeddings obtain in general lower performance scores than those with GloVe pre-trained embeddings, which might indicate that the articles on which the model was trained were not sufficient.

7.4.1 General Analysis

With our approach, which is a siamese network coupled with multiple pre-trained BERT encoders, we were able to achieve superior results, in both metrics, compared to the best-performing baselines. Between the BERT `cased` and `uncased` models, it is worth noticing that the `uncased` models perform better than the `cased` ones, which confirms that the character morphology is not important for this task, due to the fact that the capitalization is not connected to the presence of named entities or other capitalized terms. The model `bert-base-STS` (`cased` or `uncased`) did not perform as expected. We assume that the large corpus on which it was fine-tuned does not necessarily contribute to the financial domain, and thus it decreased the performance of the system. However, the difference is not statistically significant.

[5]Fine-tuned on the STSbenchmark (semantic textual similarity benchmark) dataset
[6]We used the model pre-trained on Wikipedia 2014 and Gigaword 5 (vector size 300).

Table 7.1: Experimental results for our chosen baseline models and proposed siamese-based methods.

Model	Mean Rank	Acc
Baseline Models		
– GloVe		
Logistic Regression+GloVE	1.322	0.844
Linear SVM+GloVE	1.306	0.841
Random Forest+GloVE	1.514	0.759
Decision Tree+GloVE	1.918	0.661
– FinSim embeddings		
Logistic Regression+FinSim	1.495	0.788
Linear SVM+FinSim	1.322	0.841
Random Forest+FinSim	1.527	0.743
Decision Tree+FinSim	1.951	0.657
– TF-IDF unigram		
Logistic Regression+TF-IDF	1.776	0.657
Random Forest+TF-IDF	1.469	0.743
Decision Tree+TF-IDF	1.743	0.735
Linear SVM+TF-IDF	1.453	0.8
BERT-based siamese networks		
`bert-base-uncased`	**1.2**	**0.894**
`bert-base-cased`	1.384	0.824
`bert-large-uncased`	1.241	0.886
`bert-large-cased`	1.331	0.829
`bert-base-uncased-STS`	1.220	0.882
`bert-base-cased-STS`	1.232	0.885
– definitions		
`bert-base-uncased`+definitions	1.387	0.816
`bert-base-cased`+definitions	1.363	0.832
`bert-large-uncased`+definitions	1.363	0.848
`bert-large-cased`+definitions	1.379	0.828
`bert-base-uncased-STS`+definitions	1.346	0.844
`bert-base-cased-STS`+definitions	1.359	0.840

We also evaluated which hypernyms our proposed methods failed to detect accurately. Table 7.2 illustrates the performance, using precision, recall, and F-1 measure, of our best baseline (SVM using GloVE embeddings) and the siamese network with the best performing architecture, with the pre-trained `bert-base-uncased` for the encoders. Upon evaluating these metrics as well as the predictions made, we discovered that terms containing their respective hypernym get often miss-classified, with an imbalance between the precision and the recall. “MMIs”, has ex-

ceptionally poor F1 score. We suspect that since this hypernym is an acronym, its representation might be unclear, hence could appear irrelevant.

Table 7.2: Comparing baseline with the proposed system using F1 measure

Model	Hypernym	Precision	Recall	F1
SVM+GloVE				
	Bonds	0.611	0.647	0.629
	Credit Index	0.797	0.895	0.843
	Equity Index	0.964	0.973	0.968
	Forward	**1.000**	0.500	0.667
	Funds	0.750	0.375	0.500
	Future	0.800	0.800	0.800
	MMIs	0.222	0.286	0.250
	Options	0.923	0.923	0.923
	Stocks	0.500	0.200	0.286
	Swap	0.812	0.765	0.788
bert-base-uncased				
	Bonds	0.560	0.824	0.667
	Credit Index	0.902	0.807	0.852
	Equity Index	0.948	**1.000**	0.973
	Forward	**1.000**	0.667	0.800
	Funds	0.741	0.625	0.667
	Future	**1.000**	**1.000**	**1.000**
	MMIs	0.333	0.143	0.200
	Options	**1.000**	**1.000**	**1.000**
	Stocks	**1.000**	0.400	0.571
	Swap	0.737	0.824	0.778

To take a step further, we utilized the siamese-based systems with additional information about the definition of the hypernyms to add more informative features and to obtain a better distinction between them. The definition of each concept was added to the model. These definitions were extracted from the Financial Industry Business Ontology (FIBO). as shown in Figure 7.4.1.

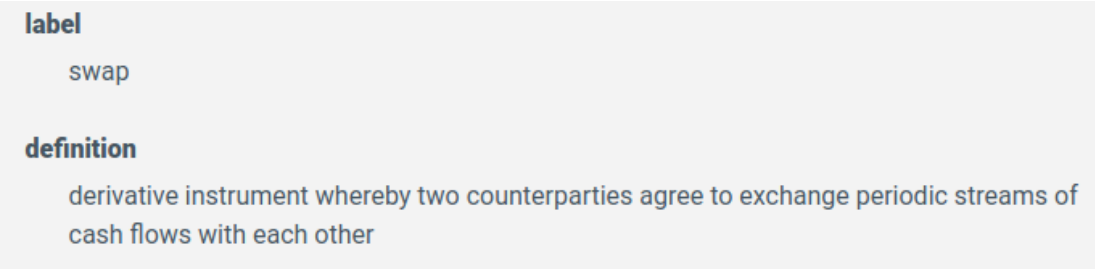


Figure 7.2: A definition from FIBO for the *Swap* ontology term.

They were concatenated with the ontology top-terms in the following manner:

(1) *Credit Default Swap.*, Swap + <hypernym definition in FIBO>, 1

(2) *Credit Default Swap.*, Credit Index + <hypernym definition in FIBO>, 0 . . .

However, the outcomes showed that adding more information will deteriorate the performance of all the siamese-based models. While it is expected that adding the definition to the label would increase the result by adding more informative features and context to the encoders, the experiment showed the opposite as both metrics slightly decreased compared to having no added information. We suspect that the definition can cause noise which affects the encoding process.

7.4.2 The Impact of the Ontology Terms

To analyze the impact of the ontology terms that can be present in the terms (cf. Table 7.2), we propose to mark the common hypernym tokens in the term (Moreno et al., 2020; Soares et al., 2019) in order to uprise their relevance. We implemented our best performing siamese-based model with BERT encoders and *EntityMarkers* (Soares et al., 2019). A BERT encoder with *EntityMarkers* consists in augmenting the input data with a series of special tokens, here named *TermMarkers*. Thus, if we consider a sentence $x = [x_0, x_1, \ldots, x_n]$ with n tokens, we augment x with two reserved word pieces (<Term> and </Term>) to mark the beginning and the end of each term in the sentence, as shown in the following example:

$Swap \subset$ *Credit Default Swap* $\implies$ (*Credit Default Swap*, Swap) $\rightarrow$ (*Credit Default* <*Term*> $Swap$ </*Term*>, Swap)

Table 7.3: The results for the best performing system with and without marked entities.

Model	Average Rank	Acc
`bert-large-uncased`	1.241	0.886
`bert-large-uncased`+*TermMarkers*	1.404	0.779

From Table 7.3, we notice that if we give more importance to the common tokens in the term and the ontology top-level term (hypernym), the results are decreasing considerably which proves that by looking at the hypernym tokens presence in the term can only diminish the performance of the system.

Table 7.4: Results of our top-3 systems on the test set provided by the organizers. Median and Best (maximum accuracy and minimum mean rank) scores are computed on the submissions from each participant, as shared by FinSim organizers.

Model	Mean Rank	Acc
L3i-LBPAM_1	1.42	0.811
L3i-LBPAM_2	1,325	0.858
L3i-LBPAM_3	1,434	0.821
Median	1.285	0.858
Best	1.189	0.906

Table 7.4 shows the results of our top-3 systems on the final test set given by the organizers. Compared to our experiments, it can be clearly seen the mean rank metric from every system is lower than expected. As for accuracy, our best results only stand among the average compared to other teams, which might hint towards the difference in the test set and the provided training sets distributions.

7.5 Conclusions

In this chapter, we tackled the task as provided by FinSim-2 the Shared Task on *Learning Semantic Similarities for the Financial Domain* by using a siamese network with BERT encoders to compute a similarity score between terms and hypernyms in a ranked manner. Our preliminary experimental results clearly outperformed the baseline, but only ranked around the median in the official scoring.

For future improvement, since the dataset was rather small, we plan to approach the task with few-shot learning. Moreover, due to the potential that this type of method could have, another aspect we need to work on is improving the model by focusing the attention mechanism on the contextual words in order to better disambiguate between the ontology and the terms.

CHAPTER 8

Conclusions and Future Work

The thesis primarily focuses on the task of Emerging Trend Detection on a variety of domain-specific datasets with a twofold objective : (I) creating robust methods that can be adapted into different types of data sources and (II) helping the “Smart Beta" team of LBP-AM in their operation of assessing potential themes for investment. To achieve these objectives, we established four questions to address the challenges we faced in our work and support our research direction.

In this chapter, we first provide our answers and findings to the research questions presented in Chapter 1. Furthermore, as we have reached the end of the thesis, we wish to present “Remarks and Perspectives” regarding the research as a whole from both academia and industry. Lastly, due to the time constraint of the thesis and limitations in various aspects such as lack of annotated datasets, evaluation framework, the scope of the data, etc., our work is left with a number of areas to develop further. Hence, we will discuss possible improvements and future directions in the last section entitled “Future Work".

8.1 Conclusions to Research Questions

At the end of our doctoral work, this section revisits its initial research questions, posed in Chapter 1.

8.1.1 Variation of the trend definition and diversity of data

Surprisingly, there is not much research regarding how the definition of a trend can vary due to the diversity in domain-specific data which can lead to the task of Emerging Trend Detection being different based on just the domain of the data. Thus, we stratified the state of the art into different categories based on the methods used and the text density in the data. Firstly, we discussed two main types of features that ETD research relies on: endogenous features that are presented in the context of the text and exogenous features that exist in external resources. Next, we present four different types of approaches in ETD based on: statistics, graphs, topic modeling, and clustering. Lastly, the design of ETD methods can depend on whether the text is dense or sparse in terms of character/word count or, simply put, text length.

8.1.2 To what extent do noise and data pre-processing impact the task of ETD?

After experimenting on multiple datasets of different domains, we draw the following conclusion to this question: the need for pre-processing a dataset is wildly different from one dataset to another, as some datasets can readily be used without any cleaning while others require the removal of undesirable parts of the text to avoid noise impacting the result of the task. Moreover, depending on the general subject, theme or context of the corpus, the process of cleaning text data cannot be directly transferred from one dataset to another due to the discrepancy in the vocabulary of domain-specific datasets where words that are considered rare in a particular corpus can be common in a specialized corpus.

8.1.3 How can context evolution of keywords help discover emerging trends?

This question is the main concern of the thesis. We conducted two experiments that utilized the surrounding context of keywords to identify potential emerging trends. First, we adopted an existing approach, Leap2Trend, by using contextualized embeddings for better representing and comparing the context evolution ranking of selected terms in a financial news setting. The second experiment, done on a corpus of computer science scientific publications, concentrated on comparing the representation of the current time interval against multiple, continuous past points to formulate a series that characterizes the context evolution of terms, which can be used as a basis for forecasting emerging trends.

We discovered that change in the context of keywords can be useful for detecting emerging trends in different scenarios. When comparing the relationship between a pair of keywords, those that have their context become more similar to each other compared to the past can signify a topic emergence, as it is the process of formulating new terms that appear in more prevalent settings. In the case of evaluating a single keyword, terms that are used in a more different context in the present are more likely to become trends as the phenomenon/situation can be interpreted that their stature and impact have branched out of its original topics and influenced other themes, which can be a good indication of emerging trends.

8.1.4 Is our approach viable in practical settings?

We apply the core idea of the thesis, which is to compare the similarity of surrounding context between words/terms, to the problem of hypernym extraction given in a shared task with a practical application and settings. The realistic data and scenario allowed us to analyze whether it is viable to use our approach to examine the closeness of words, as the objective of the task is to find words that associate with a higher-level concept. While the results in the final test set are lower than in our internal evaluation, the performance metrics we achieved using the approach were respectable compared to other submitted approaches. Thus, we believe that our method should be sufficient in a practical setting, as the accuracy in detecting lower-level concepts related to abstract level ones was more than 80%.

8.2 Remarks and Perspectives

Firstly, based on the scope of the task, the variety of domain-specific data that are available with the lack of specific annotated data for ETD, our study on state of the art and the experiment regarding noise and pre-processing, it is clear that the task was only defined vaguely at the early stage of the thesis. We viewed annotated datasets as what clearly describes the objectives of the task. However, while there is an abundance of data (in the form of Bloomberg's EDF collection) for us to research and develop our method, the absence of an ETD-specific annotated dataset in the field of finance hindered us in the evaluation. Thus, instead of just utilizing the EDF data, we opted for a number of different datasets with a wide array of problems, instead of just focusing on one specific task.

Secondly, one aspect of our experiments that we would like to reiterate is the significance of high recall in evaluation. While it is important to have a good balance between precision and recall in detection, having a high recall value is what we perceive as being more adequate since in our specific scope of research, it is the expert who ultimately decides to invest resources on particular themes. Hence, it is crucial and more favorable to have many true positives in the final result for the expert to take into consideration. In a way, this makes our task similar to that of a recommendation system.

Thirdly, and consequently, the experts at LBP-AM were given the outcome of our experiments to validate the emerging trend prediction. With prior experience and knowledge of specific domains, the experts were able to identify and confirm a few trends out of the final results when reflecting on the actual timeline and events that had happened in the time period presented in the dataset, without us showing evidence to explain the results beforehand. In the study about Microsoft, "remote work", "Microsoft team", and "digital transformation" caught the attention of the experts since they were either popular topics in the time span or in their to-observe list. Moving to less recent data with the experiment on TRENDNert, "neural network" was the most outstanding term to the "Smart Beta" team, along with "virtual machine" to a lesser extent. Moreover, trends that occurred outside these time intervals were also pointed out, such as "object recognition" and "face detection", thus giving us useful insight into the capability of our proposed systems.

Nevertheless, one concern was the black-box aspect of our approach. Indeed, a theoretical explanation of the results could be provided internally to the "Smart-Beta team" at LBP-AM through performance metrics and graphs illustrating the context evolution of a keyword in order to help demonstrate the main idea and the capability of the system. However, explaining why a term can be considered as emerging cannot be done transparently to the general audience. Experts, clients and investors from other departments and sectors do not have the essential technical knowledge to fully understand the approach. Thus, the lack of a clear visualization can hinder the persuasiveness of our approach when trying to present the results to clients and investors.

8.3 Future Work

Due to the time constraint of the thesis, we restricted some areas of the research, thus there are a number of aspects to improve upon in the future. Firstly, the data used in our experiments was limited to the subject of technology, with Microsoft-related news in Chapter 5 and computer science publications in Chapter 6. Because of this, expanding the research in order to target multiple sectors/fields of study is the main priority for improvements, as it is important for the system to be able to cover multiple domains, with a reasonable performance, in order to match the interest of LBP-AM.

Moreover, we are also curious to explore different types and sources of data aside from news and scientific publications. Social media is one of the main targets, given the influence it has over not only the public but also businesses and the stock markets. However, the main concern for using this type of data is the potential threat of fake news, as mentioned in Chapter 1, which was why we decided not to consider this data type in the thesis.

The second desired improvement for our research on ETD is to explore trends in different granularities. As our experiments showed, the main text unit used for assessing trends in our approaches is keywords, more specifically bi-grams. As a result, the keywords we assessed for trend were separated and did not have any relationship with other terms to fully connect to topics of interest. Having the capability to detect trends on multiple granularities, from abstract terms to specific concepts, and to link them together to form major emerging topics will aid in

expanding potential opportunities for considering precise themes for investment.

As we mentioned throughout the thesis, the lack of annotated data specifically designed for ETD was a hindrance to our research due to the lack of a definite ground truth and evaluation method to effectively assess the performance of our approaches. Because of this, another potential major upcoming task is to build a fully annotated dataset for the task of ETD, complete with topic labelling for each document in the corpus, marking rise/fall/fluctuation in trend movement for each topic in each time interval, and determining the period where a trend is in its emerging phase.

Lastly, we intend to tackle the concern of the experts from LBP-AM addressed in Section 8.2 about the interpretability and explainability of our results. As far as we know, people from different backgrounds will likely expect different explanations in order to be convinced and accept suggested topics as emerging trends. Thus, the first part of solving this problem is to conduct a survey of potential audiences for their opinions on what type of explanation would persuade them to believe a topic has the potential to become a trend.

Publications

This doctoral work has led to the following publications:

Conferences and workshops

- Nhu Khoa Nguyen, Emanuela Boros, Gaël Lejeune, Antoine Doucet and Thierry Delahaut: "Contextualizing Emerging Trends in Financial News Articles". Proceedings of the 4th Workshop on Financial Technology and Natural Language Processing (FinNLP) co-located with the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, December 7-11, 2022.
- Nhu Khoa Nguyen, Emanuela Boros, Gaël Lejeune, Antoine Doucet, Thierry Delahaut: "Utilizing Keywords Evolution in Context for Emerging Trend Detection in Scientific Publications". Proceedings of the 11th International Symposium on Information and Communication Technology, SoICT 2022, Hanoi, Vietnam, December 1-3, ACM, pp. 247–253, 2022.
- Nhu Khoa Nguyen, Emanuela Boros, Gaël Lejeune, Antoine Doucet, Thierry Delahaut: “L3i_LBPAM at the FinSim-2 task: Learning Financial Semantic Similarities with Siamese Transformers”. Proceedings of the Companion of The Web Conference 2021 (WWW 2021), Virtual Event, Ljubljana, Slovenia, April 19-23, ACM / IW3C2, pp. 302–306,

2021.

- Nhu Khoa Nguyen, Emanuela Boros, Gaël Lejeune, Antoine Doucet: “Impact Analysis of Document Digitization on Event Extraction”. In Proceedings of the 4th Workshop on Natural Language for Artificial Intelligence (NL4AI 2020) co-located with the 19th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2020), Anywhere, November 25-27. Vol. 2735. CEUR Workshop Proceedings, pp. 17–28, 2020.

Journal

- Emanuela Boros, Nhu Khoa Nguyen, Gaël Lejeune and Antoine Doucet: "Assessing the impact of OCR noise on multilingual event detection over digitised documents". International Journal on Digital Libraries. 23(3): 241-266, 2022.

Bibliography

Aggarwal, C. C. and Zhai, C. (2012). “A Survey of Text Clustering Algorithms”. In: *Mining Text Data* (cited on p. 25).

Akrouchi, M. E., Benbrahim, H., and Kassou, I. (2021). “End-to-end LDA-based automatic weak signal detection in web news”. *Knowl. Based Syst.*, 212, p. 106650 (cited on pp. 4, 28, 77–79).

Al-Attar, F. and Shaalan, K. (2021). “Emerging Research Topic Detection Using Filtered-LDA”. *AI*, 2, pp. 578–599 (cited on p. 20).

Alghamdi, R. and Alfalqi, K. (2015). “A Survey of Topic Modeling in Text Mining”. *International Journal of Advanced Computer Science and Applications*, 6 (cited on p. 27).

Barbaresi, A. and Lejeune, G. (2020). “Out-of-the-Box and into the Ditch? Multilingual Evaluation of Generic Text Extraction Tools”. In: *Proceedings of the 12th Web as Corpus Workshop*. Marseille, France: European Language Resources Association, pp. 5–13 (cited on p. 43).

Behpour, S., Mohammadi, M., Albert, M. V., Alam, Z. S., Wang, L., and Xiao, T. (2021a). “Automatic trend detection: Time-biased document clustering”. *Knowledge-Based Systems*, 220, p. 106907 (cited on pp. 3, 18, 27, 28, 31, 62, 78).

Behpour, S., Mohammadi, M., Albert, M. V., Alam, Z. S., Wang, L., and Xiao, T. (2021b). "Automatic trend detection: Time-biased document clustering". *Knowledge-Based Systems*, 220, p. 106907 (cited on p. 15).

Beltagy, I., Cohan, A., and Lo, K. (2019). "SciBERT: Pretrained Contextualized Embeddings for Scientific Text". *CoRR*, abs/1903.10676 (cited on p. 83).

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent dirichlet allocation". 3, pp. 993–1022 (cited on p. 18).

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). *Enriching Word Vectors with Subword Information* (cited on p. 18).

Bolelli, L., Ertekin, Ş., and Giles, C. L. (2009a). "Topic and trend detection in text collections using latent dirichlet allocation". In: *European conference on information retrieval*. Springer, pp. 776–780 (cited on pp. 15, 78).

Bolelli, L., Ertekin, S., and Giles, C. L. (2009b). "Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation". In: *European Conference on Information Retrieval* (cited on p. 62).

Boros, E., Nguyen, N. K., Lejeune, G., and Doucet, A. (2022). "Assessing the impact of OCR noise on multilingual event detection over digitised documents". *Int. J. Digit. Libr.*, 23(3), pp. 241–266 (cited on p. 44).

Boros, E., Pontes, E. L., Cabrera-Diego, L. A., Hamdi, A., Moreno, J., Sidère, N., and Doucet, A. (2020). "Robust named entity recognition and linking on historical multilingual documents". In: *Conference and Labs of the Evaluation Forum (CLEF 2020)*. Vol. 2696. Paper 171. CEUR-WS Working Notes, pp. 1–17 (cited on p. 34).

Boyack, K. W. and Klavans, R. (2022). "An improved practical approach to forecasting exceptional growth in research". *Quantitative Science Studies*, pp. 1–22 (cited on p. 3).

Brixtel, R., Lejeune, G., Doucet, A., and Lucas, N. (2013). "Any language early detection of epidemic diseases from web news streams". In: *2013 IEEE International Conference on Healthcare Informatics*. IEEE, pp. 159–168 (cited on p. 44).

Cataldi, M., Caro, L. D., and Schifanella, C. (2014). "Personalized Emerging Topic Detection Based on a Term Aging Model". *ACM Trans. Intell. Syst. Technol.*, 5(1) (cited on p. 1).

Cataldi, M., Di Caro, L., and Schifanella, C. (2010). "Emerging topic detection on twitter based on temporal and social terms evaluation". In: *Proceedings of the tenth international workshop on multimedia data mining*, pp. 1–10 (cited on p. 3).

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). *Universal Sentence Encoder* (cited on p. 27).

Chang, C.-H., Monselise, M., and Yang, C. (2021). "What Are People Concerned About During the Pandemic? Detecting Evolving Topics about COVID-19 from Twitter". *Journal of Healthcare Informatics Research*, pp. 1 –28 (cited on p. 2).

Chen, C. C., Chen, Y.-T., and Chen, M. C. (2007a). "An Aging Theory for Event Life-Cycle Modeling". *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 37, pp. 237–248 (cited on p. 25).

Chen, E. S., Stetson, P. D., Lussier, Y. A., Markatou, M., Hripcsak, G., and Friedman, C. (2007b). "Detection of Practice Pattern Trends through Natural Language Processing of Clinical Narratives and Biomedical Literature". *AMIA ... Annual Symposium proceedings. AMIA Symposium*, pp. 120–4 (cited on p. 4).

Chen, E., Lerman, K., and Ferrara, E. (2020). "Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set". *JMIR Public Health and Surveillance*, 6 (cited on p. 2).

Chen, K.-Y., Luesukprasert, L., and Timothy Chou, S. cho (2007c). "Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling". *IEEE Transactions on Knowledge and Data Engineering*, 19 (cited on p. 25).

Conneau, A. and Kiela, D. (2018). "SentEval: An Evaluation Toolkit for Universal Sentence Representations". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA) (cited on p. 93).

Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992). "Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections". In: *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '92. New York, NY, USA: Association for Computing Machinery, 318–329 (cited on p. 26).

Dang, Q., Gao, F., and Zhou, Y. (2016). "Early Detection Method for Emerging Topics Based on Dynamic Bayesian Networks in Micro-Blogging Networks". *Expert Syst. Appl.*, 57(C), 285–295 (cited on pp. 3, 20, 24, 78).

Daud, A., Abbas, F., Amjad, T., Alshdadi, A. A., and Alowibdi, J. S. (2021). "Finding rising stars through hot topics detection". *Future Gener. Comput. Syst.*, 115, pp. 798–813 (cited on pp. 21, 62, 77, 78).

De Santis, E., Martino, A., and Rizzi, A. (2020). "An Infoveillance System for Detecting and Tracking Relevant Topics From Italian Tweets During the COVID-19 Event". *IEEE Access*, 8, pp. 132527–132538 (cited on p. 2).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". *arXiv preprint arXiv:1810.04805* (cited on pp. 92, 93).

Dridi, A., Gaber, M., Azad, R. M. A., and Bhogal, J. (2019). "Leap2Trend: A Temporal Word Embedding Approach for Instant Detection of Emerging Scientific Trends". *IEEE Access*, 7, pp. 176414–176428 (cited on pp. 29, 61, 63, 68, 77, 79, 85).

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *KDD* (cited on pp. 24, 25).

Fang, Y., Zhang, H., Ye, Y., and Li, X. (2014). "Detecting hot topics from Twitter: A multiview approach". *Journal of Information Science*, 40, pp. 578 –593 (cited on pp. 21, 31).

Farzindar, A. and Khreich, W. (2015). "A Survey of Techniques for Event Detection in Twitter". *Computational Intelligence*, 31, pp. 132 –164 (cited on p. 20).

Fiscus, J., Doddington, G., Garofolo, J., and Martin, A. (1999). "NIST's 1998 Topic Detection and Tracking evaluation (TDT2)", pp. 19–24 (cited on p. 2).

Fiscus, J. G. and Doddington, G. R. (2002). "Topic detection and tracking evaluation overview". *Topic detection and tracking*, pp. 17–31 (cited on p. 2).

Galbrun, E. and Miettinen, P. (2017). *Redescription Mining*. Springer International Publishing (cited on p. 19).

Glavaš, G. and Šnajder, J. (2014). "Event graphs for information retrieval and multi-document summarization". *Expert Systems with Applications*, 41(15), pp. 6904–6916 (cited on p. 23).

Goldberg, Y. and Hirst, G. (2017). *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers (cited on pp. 17, 18).

Goodman, J., Vlachos, A., and Naradowsky, J. (2016). "Noise reduction and targeted exploration in imitation learning for Abstract Meaning Representation parsing". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1–11 (cited on p. 45).

Goorha, S. and Ungar, L. (2010). "Discovery of significant emerging trends". *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (cited on pp. 1, 26).

Grefenstette, G. (2015). "INRIASAC: Simple Hypernym Extraction Methods". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, pp. 911–914 (cited on p. 92).

Griol-Barres, I., Milla, S., Cebrián, A., Fan, H., and Millet, J. (2020). "Detecting Weak Signals of the Future: A System Implementation Based on Text Mining and Natural Language Processing". *Sustainability*, 12(19) (cited on pp. 22, 62, 77, 78).

Guha, S., Meyerson, A., Mishra, N., Motwani, R., and O'Callaghan, L. (2003). "Clustering Data Streams: Theory and Practice". *IEEE Trans. Knowl. Data Eng.*, 15, pp. 515–528 (cited on p. 26).

Gupta, P. and Jaggi, M. (2021). *Obtaining Better Static Word Embeddings Using Contextual Embedding Models* (cited on p. 19).

Hamborg, F., Lachnit, S., Schubotz, M., Hepp, T., and Gipp, B. (2018a). "Giveme5W: main event retrieval from news articles by extraction of the five journalistic W questions". In: *International conference on information*. Springer, pp. 356–366 (cited on p. 15).

– (2018b). "Giveme5W: Main Event Retrieval from News Articles by Extraction of the Five Journalistic W Questions". In: (cited on p. 46).

Hamdi, A., Jean-Caurant, A., Sidere, N., Coustaty, M., and Doucet, A. (2019). "An Analysis of the Performance of Named Entity Recognition over OCRed Documents". In: *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, pp. 333–334 (cited on p. 34).

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1489–1501 (cited on pp. 7, 10).

He, J. and Chen, C. (2018). "Predictive Effects of Novelty Measured by Temporal Embeddings on the Growth of Scientific Literature". *ArXiv*, abs/1801.09121 (cited on pp. 3, 77, 79).

He, L., Du, Y., and Zhang, L. (2018). "Vector Representation of Words for Detecting Topic Trends over Short Texts". In: (cited on p. 30).

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., and Giles, C. L. (2009). "Detecting topic evolution in scientific literature: how can citations help?" *Proceedings of the 18th ACM conference on Information and knowledge management* (cited on pp. 62, 77, 78).

Hughes, J., Aycock, S., Caines, A., Buttery, P., and Hutchings, A. (2020). "Detecting Trending Terms in Cybersecurity Forum Discussions". In: *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*. Online: Association for Computational Linguistics, pp. 107–115 (cited on pp. 20, 22, 62, 78).

Husemann, S. and Fischer, F. (2015). "Content analysis of press coverage during the H1N1 influenza pandemic in Germany 2009–2010". *BMC public health*, 15, p. 386 (cited on p. 2).

Insaf, A., Ouahabi, A., Benzaoui, A., and ahmed, A. taleb (2020). "Past, Present, and Future of Face Recognition: A Review" (cited on p. 89).

Jiao, Q. and Zhang, S. (2021). "A brief survey of word embedding and its recent development". In: *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. Vol. 5. IEEE, pp. 1697–1701 (cited on p. 18).

Jin, O., Liu, N. N., Zhao, K., Yu, Y., and Yang, Q. (2011). "Transferring topical knowledge from auxiliary long texts for short text clustering". In: *CIKM '11* (cited on p. 30).

Jones, K. S. (2004). "A statistical interpretation of term specificity and its application in retrieval". *J. Documentation*, 60, pp. 493–502 (cited on pp. 17, 21, 31).

Journet, N., Visani, M., Mansencal, B., Van-Cuong, K., and Billy, A. (2017). "DocCreator: A New Software for Creating Synthetic Ground-Truthed Document Images". *Journal of imaging*, 3(4), p. 62 (cited on p. 47).

Kassab, L., Kryshchenko, A., Lyu, H., Molitor, D., Needell, D., and Rebrova, E. (2020). "On Nonnegative Matrix and Tensor Decompositions for COVID-19 Twitter Dynamics". *ArXiv*, abs/2010.01600 (cited on p. 63).

Kaur, R. and Kaur, A. (2014). *A Review Paper on Evolution of Cloud Computing, its Approaches and Comparison with Grid Computing* (cited on p. 2).

Knobloch-Westerwick, S., Sharma, N., Hansen, D., and Alter, S. (2005). "Impact of Popularity Indications on Readers' Selective Exposure to Online News". *Journal of Broadcasting & Electronic Media - J BROADCAST ELECTRON MEDIA*, 49, pp. 296–313 (cited on p. 20).

Kohl, K. T., Jones, D. A., and Berwick, R. C. (2003). "Wordnet. an Electronic Lexical Database. Edited by Christiane Fellbaum, with a Preface By". In: (cited on p. 21).

Kontostathis, A., Galitsky, L. M., Pottenger, W. M., Roy, S., and Phelps, D. J. (2004). "A Survey of Emerging Trend Detection in Textual Data Mining". In: (cited on pp. 3, 15).

Koudoro-Parfait, C., Lejeune, G., and Roe, G. (2021). "Spatial Named Entity Recognition in Literary Texts: What is the Influence of OCR Noise?" In: *Proceedings of the 5th ACM SIGSPA-*

TIAL International Workshop on Geospatial Humanities. GeoHumanities'21. New York, NY, USA: Association for Computing Machinery, 13–21 (cited on p. 34).

Krouska, A., Troussas, C., and Virvou, M. (2016). "The effect of preprocessing techniques on Twitter sentiment analysis". In: *2016 7th international conference on information, intelligence, systems & applications (IISA)*. IEEE, pp. 1–5 (cited on p. 34).

Lee, W. S. and Sohn, S. Y. (2017). "Identifying Emerging Trends of Financial Business Method Patents". *Sustainability*, 9(9) (cited on pp. 27, 62, 78).

Lejeune, G., Brixtel, R., Doucet, A., and Lucas, N. (2015). "Multilingual Event Extraction for Epidemic Detection". *Artificial intelligence in medicine*, 65 (cited on pp. 29, 35, 44–46, 49).

Lejeune, G., Doucet, A., Yangarber, R., and Lucas, N. (2010). "Filtering news for epidemic surveillance: towards processing more languages with fewer resources". In: *Proceedings of the 4th Workshop on Cross Lingual Information Access*, pp. 3–10 (cited on p. 35).

Lejeune, G. and Zhu, L. (2018). "A New Proposal for Evaluating Web Page Cleaning Tools". *Computacion y Sistemas*, 22(4), pp. 1249–1258 (cited on p. 54).

Leskovec, J., Backstrom, L., and Kleinberg, J. M. (2009). "Meme-tracking and the dynamics of the news cycle". In: *KDD* (cited on p. 23).

Li, C., Liu, M., Cai, J., Yu, Y., and Wang, H. (2020a). "Topic Detection and Tracking Based on Windowed DBSCAN and Parallel KNN". *IEEE Access* (cited on pp. 26, 78, 79).

Li, X., Zhou, M., Wu, J., Yuan, A., Wu, F., and Li, J. (2020b). "Analyzing COVID-19 on Online Social Media: Trends, Sentiments and Emotions". *ArXiv*, abs/2005.14464 (cited on p. 2).

Lin, T., Tian, W., Mei, Q., and Cheng, H. (2014). "The dual-sparse topic model: mining focused topics and focused terms in short text". *Proceedings of the 23rd international conference on World wide web* (cited on pp. 29, 30).

Linger, M. and Hajaiej, M. (2020). "Batch Clustering for Multilingual News Streaming". *arXiv preprint arXiv:2004.08123* (cited on pp. 26, 31, 62, 77, 79).

Liu, B., Han, F. X., Niu, D., Kong, L., Lai, K., and Xu, Y. (2020). "Story Forest: Extracting Events and Telling Stories from Breaking News". *ACM Trans. Knowl. Discov. Data*, 14(3) (cited on pp. 23, 62, 78).

Liu, J., Kong, X., Xia, F., Bai, X., Wang, L., Qing, Q., and Lee, I. (2018). "Artificial Intelligence in the 21st Century". *IEEE Access*, PP, pp. 1–1 (cited on p. 88).

Lucas, N. (2009). "Modélisation différentielle du texte, de la linguistique aux algorithmes". PhD thesis. Université de Caen (cited on pp. 15, 45, 46).

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations". In: (cited on pp. 21, 25).

Majdabadi, Z., Sabeti, B., Golazizian, P., Asli, S. A. A., Momenzadeh, O., and Fahmi, R. (2020). "Twitter Trend Extraction: A Graph-based Approach for Tweet and Hashtag Ranking, Utilizing No-Hashtag Tweets". In: *LREC* (cited on p. 24).

Mansar, Y., Kang, J., and Maarouf, I. E. (2021). "The FinSim-2 2021 Shared Task: Learning Semantic Similarities for the Financial Domain". In: *Companion Proceedings of the Web Conference 2021*. WWW '21. New York, NY, USA: Association for Computing Machinery, 288–292 (cited on pp. 13, 92).

McInnes, L., Healy, J., and Astels, S. (2017). "hdbscan: Hierarchical density based clustering". *J. Open Source Softw.*, 2, p. 205 (cited on p. 25).

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). "Efficient estimation of word representations in vector space". In: *International Conference on Learning Representations (ICLR 2013), workshop track* (cited on p. 18).

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. (2010). "Recurrent neural network based language model." In: *Interspeech*. Vol. 2, p. 3 (cited on p. 61).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). "Distributed representations of words and phrases and their compositionality". In: *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pp. 3111–3119 (cited on pp. 18, 63, 92).

Miller, G. A. and Charles, W. G. (1991). "Contextual correlates of semantic similarity". *Language and Cognitive Processes*, 6, pp. 1–28 (cited on p. 18).

Moiseeva, A. and Schütze, H. (2020). "TRENDNERT: A Benchmark for Trend and Downtrend Detection in a Scientific Domain". In: *AAAI* (cited on pp. 3, 13, 38).

Monselise, M., Chang, C.-H., Ferreira, G., Yang, R., Yang, C. C., et al. (2021). "Topics and sentiments of public concerns regarding COVID-19 vaccines: social media trend analysis". *Journal of Medical Internet Research*, 23(10), e30765 (cited on p. 15).

Moreno, J. G., Boros, E., and Doucet, A. (2020). "TLR at the NTCIR-15 FinNum-2 Task: Improving Text Classifiers for Numeral Attachment in Financial Social Data". In: *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*. Tokyo Japan: -, pp. 8–11 (cited on p. 99).

Murphy, K. (2002). "Dynamic Bayesian Networks: Representation, Inference and Learning". PhD thesis (cited on p. 24).

Mutuvi, S., Doucet, A., Lejeune, G., and Odeo, M. (2020). "A dataset for multi-lingual epidemiological event extraction". In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4139–4144 (cited on p. 44).

Naaman, M., Becker, H., and Gravano, L. (2011). "Hip and trendy: Characterizing emerging trends on Twitter". *J. Assoc. Inf. Sci. Technol.*, 62, pp. 902–918 (cited on pp. 1–3, 30).

Naseem, U., Razzak, I., Khan, S. K., and Prasad, M. (2020). "A Comprehensive Survey on Word Representation Models: From Classical to State-Of-The-Art Word Representation Language Models" (cited on p. 17).

Navas Khan, R. (2020). "A Review Paper on Cloud Computing" (cited on p. 2).

Nguyen, N. K., Boros, E., Lejeune, G., and Doucet, A. (2020). "Impact Analysis of Document Digitization on Event Extraction". In: *Proceedings of the 4th Workshop on Natural Language for Artificial Intelligence (NL4AI 2020) co-located with the 19th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2020), Anywhere, November 25th-*

27th, 2020. Ed. by P. Basile, V. Basile, D. Croce, and E. Cabrio. Vol. 2735. CEUR Workshop Proceedings. CEUR-WS.org, pp. 17–28 (cited on p. 44).

Nguyen, N. K., Boros, E., Lejeune, G., Doucet, A., and Delahaut, T. (2021). "L3i_LBPAM at the FinSim-2 task: Learning Financial Semantic Similarities with Siamese Transformers". In: *Companion of The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*. Ed. by J. Leskovec, M. Grobelnik, M. Najork, J. Tang, and L. Zia. ACM / IW3C2, pp. 302–306 (cited on p. 92).

– (2022a). "Contextualizing Emerging Trends in Financial News Articles". In: *The 4th Workshop on Financial Technology and Natural Language Processing (FinNLP) co-located with the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, December 7-11, 2022*. ACM, pp. 247–253 (cited on p. 62).

– (2022b). "Utilizing Keywords Evolution in Context for Emerging Trend Detection in Scientific Publications". In: *The 11th International Symposium on Information and Communication Technology, SoICT 2022, Hanoi, Vietnam, December 1-3, 2022*. ACM, pp. 247–253 (cited on p. 78).

Nie, B. and Sun, S. (2017). "Using Text Mining Techniques to Identify Research Trends: A Case Study of Design Research". *Applied Sciences*, 7, p. 401 (cited on pp. 62, 77, 78).

Ohsawa, Y., Benson, N. E., and Yachida, M. (1998). "KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor". *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries -ADL'98-*, pp. 12–18 (cited on p. 23).

Pal, R., Chopra, H., Awasthi, R., Bandhey, H., Nagori, A., Gulati, A., Kumaraguru, P., and Sethi, T. (2021). "Predicting Emerging Themes in Rapidly Expanding COVID-19 Literature with Dynamic Word Embedding Networks and Machine Learning". In: *medRxiv* (cited on pp. 25, 63, 78, 79).

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: Global Vectors for Word Representation." In: *EMNLP*. Vol. 14, pp. 1532–1543 (cited on pp. 18, 92, 95).

Pestryakova, S., Vollmers, D., Sherif, A., Heindorf, S., Saleem, M., Moussallem, D., and Ngonga Ngomo, A.-C. (2022). "CovidPubGraph: A FAIR Knowledge Graph of COVID-19 Publications". *Scientific Data*, 9, p. 389 (cited on p. 20).

Prachyachuwong, K. and Vateekul, P. (2021). "Stock Trend Prediction Using Deep Learning Approach on Technical Indicator and Industrial Specific Information". *Information*, 12(6) (cited on p. 4).

Pugachev, L. and Burtsev, M. (2021). *Short Text Clustering with Transformers* (cited on p. 30).

Qiu, J. X., Faulkner, A., and Can, A. E. (2021). *Towards Theme Detection in Personal Finance Questions* (cited on pp. 27, 31, 78).

Reimers, N. and Gurevych, I. (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". *CoRR*, abs/1908.10084 (cited on pp. 26, 93).

Rioult, F. (2017). "Fouille de données : motifs minimaux, redescription d'espace et analyse du (e-)sport". In: (cited on p. 19).

Robertson, S. E. (2004). "Understanding inverse document frequency: on theoretical arguments for IDF". *J. Documentation*, 60, pp. 503–520 (cited on pp. 17, 21).

Robertson, S. E. and Walker, S. (1994). "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval". In: *SIGIR '94* (cited on p. 17).

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1994). "Okapi at TREC-3". In: *TREC* (cited on p. 17).

Röder, M., Both, A., and Hinneburg, A. (2015). "Exploring the Space of Topic Coherence Measures". *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (cited on pp. 58, 67).

Saif, H., Fernandez, M., He, Y., and Alani, H. (2014). "On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 810–817 (cited on p. 43).

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". *arXiv preprint arXiv:1910.01108* (cited on p. 26).

Santis, E. D., Martino, A., and Rizzi, A. (2020). "An Infoveillance System for Detecting and Tracking Relevant Topics From Italian Tweets During the COVID-19 Event". *IEEE Access*, 8, pp. 132527–132538 (cited on pp. 63, 78).

Senyo, P. K., Addae, E., and Boateng, R. (2018). "Cloud computing research: A review of research themes, frameworks, methods and future research directions". *International Journal of Information Management*, 38(1), pp. 128–139 (cited on p. 2).

Smith, R. (2007). "An overview of the Tesseract OCR engine". In: *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*. Vol. 2. IEEE, pp. 629–633 (cited on p. 47).

Soares, L. B., FitzGerald, N., Ling, J., and Kwiatkowski, T. (2019). "Matching the blanks: Distributional similarity for relation learning". *arXiv preprint arXiv:1906.03158* (cited on p. 99).

Suchomel, V., Pomikálek, J., et al. (2012). "Efficient web crawling for large text corpora". In: *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pp. 39–43 (cited on p. 43).

Swaraj, K. P., Manjula, D, Adhithyan, V, Hemnath, K. B., and Manish Kumar, L (2015). "Recent Approaches for Trend Detection from Text Documents and Real Time Streams-A Study". *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)*, 3 (16) (cited on p. 18).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). "Attention is all you need". In: *Advances in neural information processing systems*, pp. 5998–6008 (cited on p. 19).

Vogt, W. P. and Johnson, B. (2011). *Dictionary of statistics & methodology: A nontechnical guide for the social sciences*. Sage (cited on p. 16).

Wang, C., He, X., and Zhou, A. (2017). "A Short Survey on Taxonomy Learning from Text Corpora: Issues, Resources and Recent Advances". In: *Proceedings of the 2017 Conference*

on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, pp. 1190–1203 (cited on p. 92).

Wang, P., Sun, R., Zhao, H., and Yu, K. (2013). "A New Word Language Model Evaluation Metric for Character Based Languages". In: *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, pp. 315–324 (cited on p. 51).

Xie, P. and Xing, E. P. (2013). "Integrating Document Clustering and Topic Modeling". *ArXiv*, abs/1309.6874 (cited on p. 19).

Xu, S., Hao, L., An, X., Yang, G., and Wang, F. (2019). "Emerging research topics detection with multiple machine learning models". *J. Informetrics*, 13 (cited on pp. 62, 77, 78).

Zhang, C., Wang, H., Cao, L., Wang, W., and Xu, F. (2015). "A Hybrid Term-Term Relations Analysis Approach for Topic Detection". *Knowledge-Based Systems*, 93 (cited on p. 23).

Zhu, Z., Liang, J., Li, D., Yu, H., and Liu, G. (2019). "Hot Topic Detection Based on a Refined TF-IDF Algorithm". *IEEE Access*, 7, pp. 26996–27007 (cited on pp. 21, 31, 62, 66, 78).

Zou, Z., Shi, Z., Guo, Y., and Ye, J. (2019). *Object Detection in 20 Years: A Survey* (cited on p. 89).

Zubiaga, A., Spina, D., Martínez-Unanue, R., and Fresno-Fernández, V. (2015). "Real-time classification of Twitter trends". *Journal of the Association for Information Science and Technology*, 66 (cited on p. 36).