Cloud Infrastructure for Intra- and Inter-Social Media Platforms News Dispersal Pattern Analysis

Jiachen Li (1068299)

Supervisors:

Prof. Richard Sinnott

Research Proposal

April 2023

Word count: 5996

Master of Computer Science

School of Computing and Information Systems

The University of Melbourne

Abstract

Social media is inevitably replacing traditional news media to become the primary source for people to obtain news and information to guide their daily decisions. The fast information flow nature of social media has made an undeniable contribution to its dominant position today. Unfortunately, this nature of social media has been a double edge sword that also allowed inaccurate or, even worse, fake information to spread and potentially harm entire society by manipulating people's critical decisions. There has been tremendous research conducted regarding automatic social media misinformation detection that the majority of models were heavily relying on social media posts' propagation-based features. However, the rumour detection approach is limited to binary output that classifies the social media post into absolute truth or fake, which is not comprehensive enough for end users due to most posts in real life being a mixture of both. Such systems are also incapable of detecting news social media posts based on truth but told in a misleading tone for ideology propaganda purposes. Therefore, it is vital to analyse the news information dispersal pattern in-depth to offer end-users more comprehensive information to think critically about how much trust should be put in each social media news post based on the action and ideologies the sender took when forwarding and potentially manipulating the news. To conduct such research, we propose to model implicit content networks to track news information propagation across different social media platforms and analysis the captured intra- and inter-platform dispersal pattern with knowledge graph-based strategies. As the misinformation problem is a large-scale social problem, effort from the field of applied computing and distributed systems is also required. In this research, we also aimed to experiment and engineer a cloud-based system that takes advantage of trending cloud technologies with the benefits of being accessible, scalable, and robust to offer adequate performance and accessibility to gather data torrents from multiple social media platforms, including Twitter, Reddit, and Mastodon to perform proposed natural language processing models.

Contents

1 Introduction

In the past few years, the rise of social media has been one of the most significant trends in the entire world. Thanks to the ease of use and rapid flow of information on social media, it has been integrated into people's daily life to connect with others. Especially in recent years, social media is not limited to being a source of entertainment but is also used to consume news as a source of information. A Survey conducted by Pew Research Center suggests that 53%, which is the majority of adults in the United States, often or sometimes get news from social media in 2020 [1], while four years ago, in 2016, only 44% of adults in the United States, often or sometimes get news from social media [2]. The survey has illustrated that social media is rapidly but ineluctably replacing traditional web-based internet news to become the primary source of news content for most people because it can offer more personalised, diverse, and interactive news content. In fact, journalists who are responsible for producing traditional news based on paper or web media rely on social media for sourcing or verifying news. In the study conducted by Xinzhi Zhang and Wenshu Li, 255 participant journalists working in Hong Kong, on average, 54.8% agree they use social media for news sourcing, and 48.8% agree they use social media for news verification [3].

However, also due to the great accessibility and fast information flow, social media poses various challenges and risks, including but not limited to misinformation, addiction, privacy concern and cyberbullying. Among all the challenges and risks, misinformation is today's society's most severe challenge. The misinformation on social media can result in grievous damage to but not limited to public health, democracy, and social cohesion, as people are overlying relying on the news and information obtained from social media to guide their daily and critical decision. The study conducted by Karlsen and Aalberg suggests that the existence of misinformation on social media may influence people's perceptions resulting in people not trusting any news, including news

from traditional paper or web-based media [4]. Besides the trust issues, the misinformation on social media may also manipulate emotions to polarise opinions, ultimately inciting violence.

In response to hazardous misinformation on social media challenges, numerous research has been conducted in the NLP field. One approach is rumour detection, which uses various features from social media posts and posting accounts to classify the post as rumour or truth at an astonishing accuracy. However, the approach has multiple limitations, such as only providing binary information that may be too absolute to ask the end user to trust or disregard a post entirely or failing to classify manipulated misleading information based on truth. Therefore, the alternative news dispersal pattern analysis approach was proposed to encounter such limitations. By studying how the news was spread and modified across different or within the same users or social media platforms, the characteristics and ideology of the user or social media platforms influencing the way forwarded news may be manipulated, which can help the reader to decide how much information they want to take from the post.

Either approach would require vast training data and exceptional computational power. To fundamentally solve the negative influence of misinformation, it is also critical to make the experimental outcome available to the researcher and the general public. The trending cloud technology satisfied the requirement of scalability and accessibility with additional robustness. Therefore, we propose to research and build a cloud-based infrastructure to analyse news dispersal patterns across social media posts collected from various social media platforms, including Twitter, Reddit and Mastodon.

2 Literature Review

This section will first explore the development of social media rumour detection research in relation to the importance of dispersal information. Then different approaches to tracking news dispersal will be explored, followed by knowledge graph-based analysis strategies. Finally, the benefits of cloud computing will be illustrated to support the necessity of using such technologies for the investigated task.

2.1 Rumour Detection

In natural language processing, social media rumour detection is an important research topic aimed at identifying the social media post containing misinformation to avoid negative influence. The rumour detection field was pioneered by Castillo et al. in 2011, who proposed an automatic method for assessing the credibility of a given set of tweets [5]. The research was conducted with data from Twitter on newsworthy topics that were getting a spike in popularity in a short time detected by Twitter Monitor [6]. J48 decision tree machine learning model was selected, shown in Figure 1, with the output of "A" for truth and "B" for the rumour. The input features were from three different aspects: user, topic, and propagation. In contrast, the message-based feature is ineffective for the task. Benefiting from the excellent interpretability of the decision tree machine learning model, we learnt that the core idea was to identify if the tweet was sent from a user in connection with a network of credible users that are more likely to produce trustworthy tweet posts. The fine-tuned model claimed to achieve an astonishing 0.787 F1 score considering it was constituted by a non-neural network model completed in the early stage of the research field. We observed that the paper has suggested the URLs in tweets were traded as an important feature under the propagation aspect feature set. This has coincided with the idea of URLs based

information propagation tracking to be reviewed in detail in the following sections.

Figure 1: Decision tree model built for the credibility classification by Castillo et al. [5]

After Castillo et al.'s pioneered paper, abundant research has been conducted to improve the performance of the automatic rumour classifier by tweaking different parts of the model's pipeline or specialised the model to perform in a specific setting.

One such research conducted by Sicilia et al. focuses specifically on the social media post data from Twitter in the health domain for rumour detection [7]. The main inno-

vation point of the paper is that Sicilia et al. proposed a novel feature design method to improve accuracy across various classification machine learning models. Because of the limitation of social media post data from only a single domain, the topic-based features mentioned in the previous study were unavailable; therefore, Sicilia et al. have proposed a range of new features and benchmarked them with other remaining available features across different machine learning classification models. As a result, the finding has been shown in Figure 2 cited from the original paper [7]. In the figure, the newly proposed influence potential-based features are marked with a dagger (†) and graph theory-inspired features are marked with a diamond (♢). The histogram can be interpreted in a way that more informative features would have a shorter stacked histogram bar, and the red dotted vertical line stands for the median. For example, the most informative feature, the sentiment score, suggests that tweets with neutral tones tend to be more trustworthy. The paper concluded that the proposed new features enable models to "detect about 90% of rumours, with acceptable levels of precision" [7]. From the result, we observed that features related to the information dispersal pattern tend to be more informative compared to the feature related to user information. And URL information played an essential role in the task, such as the feature of the likelihood of a URL being shared (Purl shown in Figure 2).

Besides improving the performance by tweaking the features, there have also been researches experiment on rumour detection with more advanced machine learning models, especially in the current era where the rising neural network-based deep learning strategies have outperformed traditional machine learning models in the vast majority of aspects. The research conducted by Tian et al. focused on the early detection of rumours on Twitter [8]. Rumour detection in the early stage is essential for the platform to contain the spread of misinformation in a timely manner to reduce the negative influence as much as possible. However, at the early stage, before the information starts

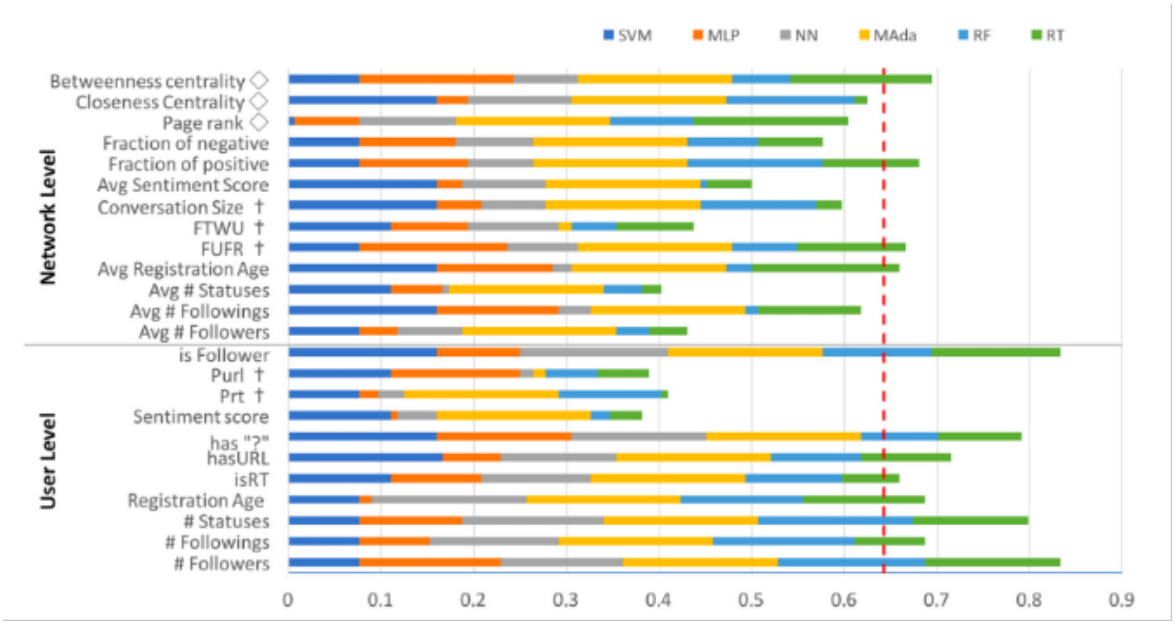

Figure 2: Stacked histogram of the rank analysis by Sicilia et al. [7]

to spread, the available features are minimal. To tackle such a challenge, the early rumour detection model proposed by Tian et al. only relied on features from tweet contents and their immediate user comments. The model consists of two procedures. The first step is to learn the attitude representation for stance prediction from user comments, and Tian et al. proposed to use CNN and BERT-based deep neural models to solve such tasks via transfer learning. The followed up step was to classify the rumours based on the attitude representation obtained from the previous step and the content representation of the tweets with their comments. For this step, the model proposed by Tian et al. was based on CNN-BiLSTM and BERT neural models. The experiment was conducted on two public Twitter datasets inspired by Ma et al. [9], namely, Twitter15 [10] and Twitter 16 [11]. Tian et al. claimed that their proposed BERT-based model had been consistently outperforming state-of-the-art baselines significantly. The research demonstrated that the user comments are informative and contain early signals for rumour detection.

The last study to be discussed was conducted by P et al. on the automatic detec-

tion of rumoured tweets and finding their origin [12]. The research was conducted on the data set containing Twitter posts relating to the 2011 London Riots with the fact check done by The Guardian. Although P et al. have used the same J48 decision tree machine learning model as Castillo et al. [5], P et al. have innovatively conducted a feature reduction based on the experiment results to improve the classification accuracy. It was found that user-propagated content-based features are considerably more important than the user's identity-based feature. This was explained by the observation illustrated by Gupta et al. [13] that the genuine user has an equal chance of retweeting and propagating tweets containing misinformation as the rogue users intend to spread rumours during the crisis situation. To better understand the potential of user-based features, P et al. proposed an algorithm for locating the origin rumour posting user for user-based features analysis to help the situation with limited tweet-based information, similar to the early stage rumour detection setting in Tian et al.'s research [8].

From the discussed social media rumour detection research, we observed that modern rumour classifiers could achieve exceptional precision. At the same time, considerable research has shown that the URL information in tweets relates to the dispersal pattern, which is a more informatic feature. Also, observing that information dispersal patterns or networked-based features are often more important than text or user statistics has emphasised that information dispersal pattern analysis is crucial. Another reason for researching news dispersal patterns is to overcome one of the main limitations of the rumour detection approach, which classifies tweets only into binary classes of truth or rumour. As the data preparation section from the discussed paper suggested, in many situations, the tweet may not be wholly truthful or totally implausible. The majority of tweets have in-between grayscale trustworthiness rather than complete white or black. In most rumour detection studies, tweets that genuinely influence the user and our society have been discarded as "noise" or "non-newsworthy" data. We propose that a more in-depth analysis of the social media posts based on their dispersal pattern will

offer end-user more comprehensive information on how much trust should be put in each tweet based on the action and viewing point the sender took in its news dispersal pattern. For example, showing the reader a specific user or platform that tends to tell more about the negative aspect of vaccine-related news would provide more information on the sender's ideologies and intentions rather than mark the news as rumour or truth, would encourage the reader to think more critically before fully trusting or ignoring the news.

There are also other limitations, such as the definition of "rumour" can be unclear or sometimes biased in many situations, namely predictive information; for example, tweets regarding the prediction of the future cannot be simply classified as truth or rumour. Also, when using rumour detection models to solve the initial problem of preventing social media news information from harming society, the rumour may not always be harmful, and more importantly, the truth may not always be safe. In fact, Some manipulated or polarised information is a fraction of the truth or a combination of multiple truths that may be classified as truth but still harmful. An example of such disingenuous news information would be the "#DirectedEvolution" video [14], widely spread in February of 2023. The original video claims that the major covid-19 vaccine producer, Pfizer, purposely directed the evolution of covid 19 on monkeys in the laboratory environment to test the effectiveness of their product against potential mutation of covid-19, which Pfizer has confirmed that it was done in a regulatory and safe way [15]. While we remain neutral regarding this news, including its facticity, it was observed that as the news is propagating on Twitter, it has been manipulated as "Pfizer was making money from the Pandemic by selling vaccine, and they have purposely directed the evolution of covid 19". Strictly speaking, the statement contains no misinformation, but it has purposely omitted the information that the experiment was conducted in a safe laboratory environment to trick readers into making the inference that Pfizer has been actively directing the evolution of covid 19 to extend the pandemic

for financial benefits. The conclusion that the sender is manipulating the information to spread anti-vaccine ideologies is more informative and helpful for researchers and users of social media. Therefore, it's necessary to study the information dispersal patterns to understand how such news flows and how fake news can be dealt with in a timely manner. Researching the news dispersal patterns can also help users identify the ideology or stand of an account or social media platform that may influence their sourced news information to encourage users to think critically, as mentioned before.

2.2 News Dispersal Pattern Tacking

There has not been previous news dispersal pattern analysis research conducted on social media news as the rise of social media as a news source was so fast and recent. There have been numerous research on news dispersal pattern analysis on traditional paper-based and web-based news. Most of such papers are not applicable to our study due to the dramatic difference in news form between conventional media and novel social media. However, exceptions do exist; research taking the approach of analysing the news at the content level would overcome the form difference and be applicable to news from social media platforms. The content-based news dispersal analysis also has another advantage of interpretability to ordinary users. Joshi and Sinnott demonstrate the limitation of existing classic commercial aggregation systems that stop at document-level clustering and frequency analysis to leave the responsibility to read the content thoroughly for complete understanding and unearth the relation between contents to make the judgment to end users [16]. This goes against the initial goal of ensuring that the system is easy to use and helps everyday users on a regular basis. Based on the two reasons illustrated above, the paper will advance the research solely with content-focused modelling approaches rather than entity tracing done by Minard et al. [17] or knowledge graph modelling done by Rospocher et al. [18]. Therefore, this

section will first review the development of content-based news propagation analysis in chronological order, then expand the discussion on the latest state-of-the-art solution proposed by Joshi and Sinnott [16].

Macroscopically speaking, the development of content-level propagation analysis can be divided into two major stages: the link diffusion method and the content diffusion method. The link diffusion system utilises URLs as a pointer to connect news into a propagation graph. The research of the link diffusion method was pioneered by Adar and Adamic using epidemiological models to generate a propagation graph based solely on links from news [19]. Later on, the content diffusion approach was introduced to overcome the limitation of the link diffusion model that only utilises links, which only make up a small portion of the news content. The early content diffusion approach proposed by Young and Leskovec exactly matches the conserved content, mainly quotes or Twitter hashtags between news, to track the news propagation [20]. On the other hand, MemeTracker System proposed by Leskovec et al. allows quotes, or "memes" in this setting, to be slightly different in each propagation [21]. The approach was further improved by Sjuen et al. in their NIFTY system to introduce a clustering algorithm to track the slight change of quotes over time [22]. However, the approach still only relies on quotes, the short text enclosed in quotes that also only make up a limited portion of full text, similar to URLs. Colavizza et al. used similarity measures and local text alignment to use the full text of the news fully [23]. On the one hand, the proposed approach has finally fully utilised the entire content of the news article for the first time; on the other hand, the approach had a new limitation of being unable to match semantically similar content as the similarity measure, string kernel in this approach, only captures the appearance of text. Thereafter, a more advanced approach to address this new issue has been proposed by Vakulenko et al., which approaches the problem by phrasing art articles into "n-gram-like" grammatical relations at the sentence level to track the dispersal of information [24]. While it has overcome the

shortcoming of previous studies, the main limitations of this model are incapable of monitoring more complicated content relationships and the dependency on the specific language grammatical parsers and the synsets in Wordnet [25].

The model proposed by Joshi and Sinnott has summed up the experience of predecessors, standing on the shoulders of giants [16]. The dataset used for training the model was The Single Media 1-Million Article Training Dataset [26] because it is a standardised large dataset obtained from real online sources that include duplication, noisy data, incorrect language articles and unparsable text that best simulated the real word situation. The duplication removal news article pre-processing was applied to avoid the output graph being too dense with similar nodes. The first step they did was to use doc2vec [27] to vectorise the paragraph, which was an extension of word2vec [28] that vectorises words into high-dimensional vectors [29] that are distributed representations of words with semantic meaning encoded [30]. Joshi and Sinnott have taken suggestions from the reference model proposed by Lau and Baldwin on using a distributed bag of words method to train the model and optimal hyper-parameter settings [31]. The vectorisation model was then evaluated by the English Semantic Textual Similarity (STS) task from *SEM [32] and SemEval [33, 34, 35, 36] to test the performance of semantic models in different language domains to measure the pairwise similarity of sentences according to Joshi and Sinnott [16]. After vectorising the sentence, the cosine similarity [37] of sentence vectors was used as a distance for the k-nearest neighbours (K-NN) [38] algorithm in conjunction with the hierarchical agglomerative clustering (HAC) [39] algorithm to create the propagation graph. Result-wise, according to STS evaluation, the proposed semantic model works exceptionally well in the news domain. For the clustering stage, HAC appeared to be essential to avoid overlap clusters generated by K-NN, even though HAC is a much more computationally demanding algorithm. Finally, the output is visualised by connecting the nearest neighbour sentence forming the news dispersal pattern graph for further analysis.

2.3 Knowledge Graph-based News Posts Analysis

After representing the pattern of news propagation as a network graph, the next step is to analyse the modification of content at each node to uncover the peculiarity of the sender or platform. The news event analysis and forecasting toolkit based on the knowledge graph proposed by Hassanzadeh et al. would be helpful for such a scenario [40]. The knowledge base was a class of strategies for storing information, knowledge, and truth in the computer system, which knowledge graph (KG) has been the most widely accepted intense of knowledge bases in the question-answering field of NLP research. The KG is constructed by Resource Description Framework (RDF) triple in format ¡subject, predicate, object¿. The toolkit proposed by Hassanzadeh et al. has three main functions: extracting the news post into KG, using KG for causal analysis and forecasting, and supplementing KG with causal knowledge from external sources. The distinctive point of Hassanzadeh et al.'s research is that their toolkit enables building a human-in-the-loop explainable solution for downstream analysis, which has met our initial accessibility goal. However, the toolkit was only capable of retrieving news headlines from Wikinews and EventRegistry [41] with no support for social media news post input.

The task of retrieving news posts into KG can be classified as a relation extraction task within the NLP field. To solve such a class of problems, Trisedya et al. have proposed a novel neural-based end-to-end-relation extraction model for information extraction [42]. Their approach has addressed the error propagation drawback caused by Named Entity Disambiguation (NED), which was heavily relied on for mapping RDF to knowledge base space, by replacing NED with a novel neural-based end-to-end model. The paper also improved the performance from other aspects. An n-gram-based attention model was proposed to avoid breaking multi-word entity names from news posts apart during the extraction.

The task of merging different representations of the same real-world entities is called Entity Alignment (EA). It is essential for creating more precise KG or connecting multiple KGs together to expand the knowledge base. To address the issue of merging similar but distinct entities using a traditional embedding similarity-based greedy search algorithm, a modified beam search along with a triple classification algorithm was proposed by Trisedyl et al. to merge the same real-world entities with different names in KG [42]. To further improve the performance, Trisedyl et al. have published a separate paper on the EA problem that first generates attribute embeddings from the attribute triples as the seed alignment to shift the entity embedding of multiple KG into the same vector space [43]. Then the number of attributes of an entity is further enriched by the method transitively to identify the similarity between different entity embeddings better.

2.4 Cloud Computing

All the previously mentioned algorithms and models are either computationally expensive or require unprecedented amounts of data. We propose to use cloud technologies to address the above issue with the extra benefit of broad network access, rapid elasticity and other essential characteristics of the cloud proposed by Mell and Grance [44]. Australian Data Observatory (ADO) [45], funded by the Australian Research Data Collection, was built to collect, store, and analyse social media posts, according to Morandini [46]. The ADO system shared various similarities compared to the cloud system we proposed, except ADO was not focused on the news alone so that only classified posts into finite topics. In contrast, the news may have new issues come up daily. By interviewing members of the Melbourne eResearch Group, mainly Morandini and Sinnott, who were responsible for developing and operating ADO, valuable advice

was learnt to aid the further development of our similar cloud system. One of the main challenges and resolvent is the data overflow issue. Starting from June 2021, ADO has started collecting social media posts related to Australia [46]; the enormous amount of data has cluttered the storage resources that the University of Melbourne can ever provide. To encounter such a challenge, Professor Richard Sinnott, the Director of eResearch Group, has innovatively decided to convert ADO from a university-owned private cloud to a hybrid cloud that bursts into a public cloud when extra storage is needed.

3 Methodology

3.1 News Dispersal Pattern Analysis

After being inspired by the research conducted by Joshi and Sinnott, we have come up with several potential approaches to enhance performance. Joshi and Sinnott's paper was published in 2018 when social media was not trending for news consumption; it was also a time when the pre-trained language models had not dominated the NLP research field. Therefore, instead of the doc2vec-based encoder [27], we propose to use the state-of-the-art Bidirectional Encoder Representation from Transformers (BERT) introduced by Devlin et al. from Google in 2019 [47]. The BERT is a multi-layer bidirectional transformer encoder-based model that is pre-trained on two enormous corpora: BooksCorpus(800M words) [48] and English Wikipedia (2,500M words). By adding an additional output layer to any existing neural architecture, the bidirectional text representation learnt by BERT can be applied to any downstream NLP task [49], including text clustering. It is also worth noting that text from social media tends to differ dramatically from traditional newspapers with less formal language and short passage length. Therefore, BERTweet, proposed by Nguyen et al., explicitly designed for English Tweets, would be a better fit [50]. Also, in order to filter the none revalued posts from abundant social media data, we proposed to use the method introduced by Morandini et al. to use the BERTopic [51] for the topic modelling task in the pre-processing stage to only preserve the posts related to an interested news topic [46]. We also propose to include the "popularity" or the importance of a specific node based on its influence in the final visualisation stage; with an additional dimension, we intend to use the topographic visualisation method proposed by Morandini et al., demoed in Figure. 3 to replace the traditional network graphs [46].

Figure 3: Topographic visualisation method on Batman Peak visualisation by Morandini et al. [46]

3.2 Cloud Computing Platform

The Melbourne Research Cloud (MRC) is chosen as the cloud platform to conduct experiments on the proposed social media news dispersal pattern analysis model. The MRC was managed by the University of Melbourne, which grew out of the Nectar Research Cloud [52], Australia's national research cloud specifically designed for research computing [53]. Under the cloud taxonomy discussed in Journal by Mell and Grance, the MRC is classified as an Infrastructure as a Service (IaaS) service model and private cloud deployment model [44]. The MRC system is built based on the standardised open-stack framework allowing access to different modules for extra functionality.

The first benefit of using MRC is its availability. As a cloud-based system, MRC instances are capable of running day and night continuously and thus enable us to run the social media post collection and analysis backstage continuously. Secondly, with allocated resources from MRC illustrated in Table 1, sufficient computational resources

have been provided to complete this research project, especially since sufficient RAM is essential for large-scale NLP models. Even though MRC provides great on-demand dynamic scalability, extra resources can always be applied and allocated in the event of a burst in the news. In the security aspect, MRC, as a private cloud, has implemented stringent security. The security group function avoids unauthorised access, and the snapshot function help system to recover from a fetal crush.

Resource	Quality
Compute Resources - Virtual Cores	32 VCPUs
Compute Resources - Instances	6 servers
Compute Resources - RAM	Unlimited
Volume Storage	2000 GiB
Advanced Networking - Networks	3 Networks
Advanced Networking - Routers	3 Routers
Advanced Networking - Floating Ips	2 Floating IPs
Advanced Networking - Load Balancers	2 Load Balancers

Table 1: MRC resource allocation

3.3 Cloud System Architecture

The MRC is classified as an Infrastructure as a Service (IaaS) delivery model according to the NIST definition of cloud computing written by Mell and Grance [44]. In this model, the cloud service providers manage the networking, storage, servers, and virtualisation to offer users a cloud infrastructure to deploy and run arbitrary software, including cooperation systems and applications. The user is responsible for managing and controlling the operating system, middleware, runtime, data, and application. IaaS is optimal for conducting this research as we have absolute control and freedom over the software side while the hardware has been encapsulated. Therefore, we have

designed a system architecture specifically for this research to utilise the full potential of the cloud computational power while being robust. The system's fundamental design is based on the master-worker architecture with a worker pool design to manage multiple instances of a distributed and parallel computing system. The master is opened to the public internet to receive news topics of interest for analysis from researchers or users through a ReSTful HTTP request from the researcher or user. The ReSTful design of the master node enables standardised access to input requests and output results for third parties to develop an open-source frontend quickly. After receiving the request, the master will decomposite the task into smaller subtasks and add them to the task queue stored in the database system. The worker instance periodically checks the task queue and take task when available. This design can bring two significant benefits, scalability and fault tolerance. Each worker is running identical code capable of processing any task, enabling the system to scale horizontally by adding or taking additional worker nodes with no extra configuration needed while the system runs and functions normally. Fault tolerance-wise, by adding the time-out logic that when work fails to return the result of a particular task due to a potential node crash or network outage, the master node will add the assigned task back to the top of the task queue.

3.4 Data Source and Storage

The developer API for Twitter [54], Reddit, and Mastodon will be used to stream news-related posts continuously to the buffer that keeps posts for seven days to avoid data overflow. When trending news is detected, the system can also use API to search history posts with keywords for up to seven days due to the limitation of API. The collected data will be stored in the database system of choice, which in this research is CouchDB [55]. Two main benefits brought by CouchDB are clustering for fault tolerance and MapReduce for efficient data preprocessing [56]. By running a CouchDB cluster across

different nodes, the replication of the database will store the same document on different nodes to reduce the chance of data loss and improve fault tolerance. The in-build MapReduce query logic can be applied for content extraction from raw data in the preprocessing stage or statistics summarisation in an easy-to-manage and efficient way as the algorithm will be running in parallel across clustered CouchDB deployed on different nodes.

3.5 Future Improvement

It is essential to mention that we are aware of the inexorable momentum of short videos taking over the internet, similar to how traditional internet-based news media was replaced by social media. It was feasible to apply speech-to-text and image-to-text algorithms to feed the outputted text as the input to discussed unsupervised learning model discussed previously in this paper. However, unfortunately, at the time of this paper being written, the leading company of the short video industry, TikTok, confirmed by its CEO, Shou Chew, that it has 150 million monthly active users in the United States, which is making up nearly half of its countries population has not been opening their researcher API access to a researcher outside of United States. Future follow-up research on including TikTok short video to the news dispersal pattern analysis will be conducted once the TikTok developer API is available globally in the near future.

4 Timeline

The planned time span of this research project is one school year which was the composition of one winter break (WB) and two school semesters of 12 weeks each. The detailed research time allocation has been shown in Figure 4. Note that the social media developer key application has been made sooner rather than later due to the limitation of seven days of search windows limitation on Twitter developer API V2.0.

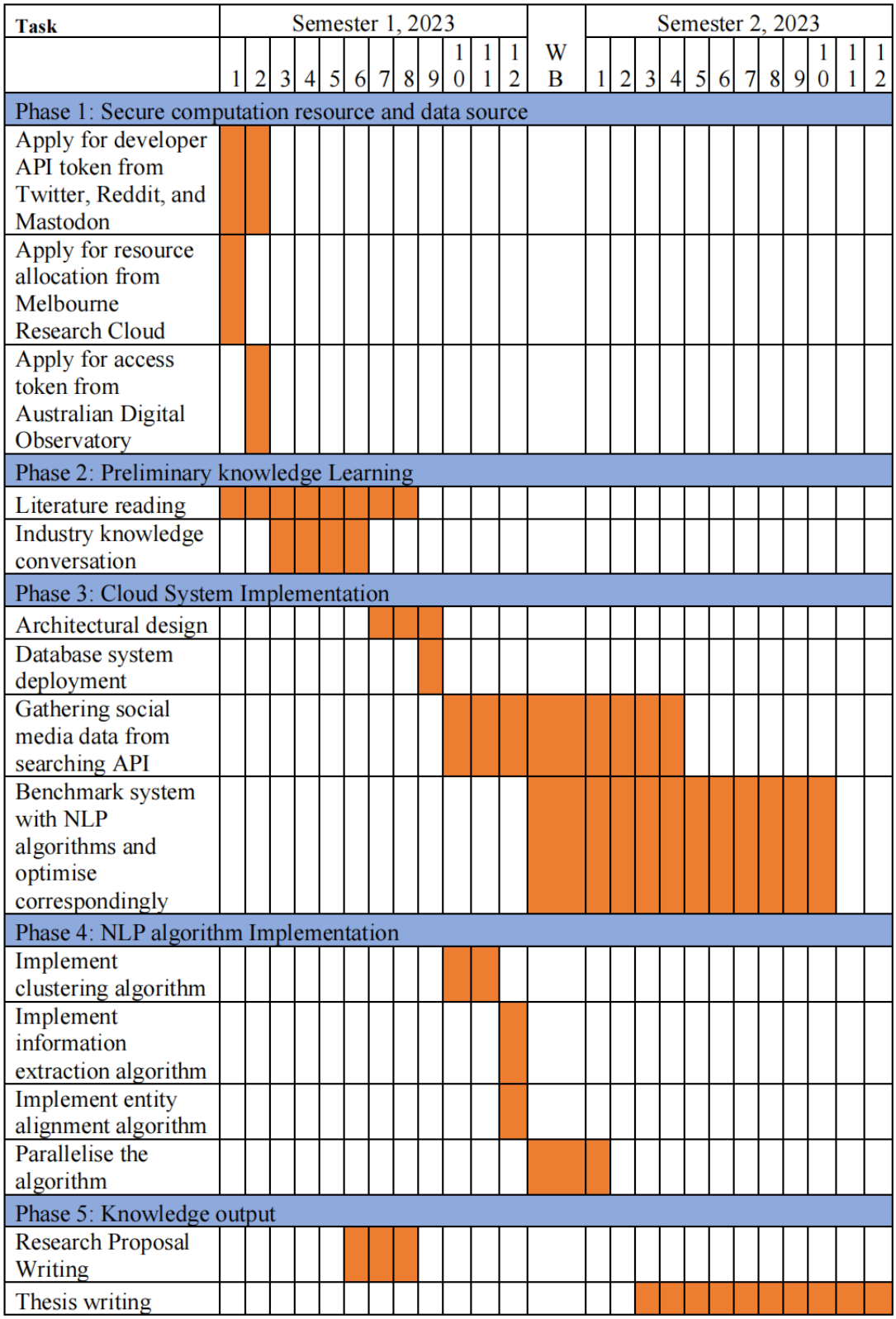

Figure 4: Research Timeline

References

[1] E. Shearer and A. Mitchell, "News Use Across Social Media Platforms in 2020." [Online]. Available: https://www.pewresearch.org/journalism/wp-content/uploads/sites/8/2021/01/PJ_2021.01.12_News-and-Social-Media_FINAL.pdf

[2] J. Gottfried and E. Shearer, "News use across social media platforms 2016." [Online]. Available: https://www.pewresearch.org/journalism/wp-content/uploads/sites/8/2016/05/PJ_2016.05.26_social-media-and-news_FINAL-1.pdf

[3] X. Zhang and W. Li, "From Social Media with News: Journalists' Social Media Use for Sourcing and Verification," vol. 14, no. 10, pp. 1193–1210. [Online]. Available: https://doi.org/10.1080/17512786.2019.1689372

[4] R. Karlsen and T. Aalberg, "Social Media and Trust in News: An Experimental Study of the Effect of Facebook on News Story Credibility," vol. 11, no. 1, pp. 144–160. [Online]. Available: https://doi.org/10.1080/21670811.2021.1945938

[5] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th International Conference on World Wide Web*, ser. WWW '11. Association for Computing Machinery, pp. 675–684. [Online]. Available: https://dl.acm.org/doi/10.1145/1963405.1963500

[6] M. Mathioudakis and N. Koudas, "TwitterMonitor: Trend detection over the twitter stream," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '10. Association for Computing Machinery, pp. 1155–1158. [Online]. Available: https://dl.acm.org/doi/10.1145/1807167.1807306

[7] R. Sicilia, S. Lo Giudice, Y. Pei, M. Pechenizkiy, and P. Soda, "Twitter rumour detection in the health domain," vol. 110, pp. 33–40. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418303129

[8] L. Tian, X. Zhang, Y. Wang, and H. Liu, "Early Detection of Rumours on Twitter via Stance Transfer Learning," in *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science, J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, and F. Martins, Eds. Springer International Publishing, pp. 575–588.

[9] J. MA, W. GAO, and K.-F. WONG, "Detect rumors in microblog posts using propagation structure via kernel learning," pp. 708–717. [Online]. Available: https://ink.library.smu.edu.sg/sis_research/4563

[10] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time Rumor Debunking on Twitter," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, ser. CIKM '15. Association for Computing Machinery, pp. 1867–1870. [Online]. Available: https://dl.acm.org/doi/10.1145/2806416.2806651

[11] J. MA, W. GAO, P. MITRA, S. KWON, B. J. JANSEN, K.-F. WONG, and M. CHA, "Detecting rumors from microblogs with recurrent neural networks," pp. 3818–3824. [Online]. Available: https://ink.library.smu.edu.sg/sis_research/4630

[12] S. V. P, A. R. Pias, R. Shastri, and S. Mandloi, "Automatic detection of rumoured tweets and finding its origin," in *2015 International Conference on Computing and Network Communications (CoCoNet)*, pp. 607–612.

[13] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy," pp. 729–736.

[14] "#DirectedEvolution TAKES THE WORLD BY STORM!!" [Online]. Available: https://www.youtube.com/watch?v=BLE_3JElSNs

[15] Pfizer Responds to Research Claims — Pfizer. [Online]. Available: https://www.pfizer.com/news/announcements/pfizer-responds-research-claims

[16] A. Joshi and R. O. Sinnott, "Modelling Implicit Content Networks to Track Information Propagation Across Media Sources to Analyze News Events," in *2018 IEEE 14th International Conference on E-Science (e-Science)*, pp. 475–485.

[17] A.-L. Minard, M. Speranza, E. Agirre, I. Aldabe, M. Van Erp, B. Magnini, G. Rigau, and R. Urizar, "SemEval-2015 Task 4: TimeLine: Cross-Document Event Ordering," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, pp. 778–786. [Online]. Available: http://aclweb.org/anthology/S15-2132

[18] M. Rospocher, p. u. family=Erp, given=Marieke, P. Vossen, A. Fokkens, I. Aldabe, G. Rigau, A. Soroa, T. Ploeger, and T. Bogaard, "Building event-centric knowledge graphs from news," vol. 37–38, pp. 132–151. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1570826815001456

[19] E. Adar and L. Adamic, "Tracking Information Epidemics in Blogspace," in *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*. IEEE, pp. 207–214. [Online]. Available: http://ieeexplore.ieee.org/document/1517844/

[20] J. Yang and J. Leskovec, "Modeling Information Diffusion in Implicit Networks," in *2010 IEEE International Conference on Data Mining*, pp. 599–608.

[21] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. Association for Computing Machinery, pp. 497–506. [Online]. Available: https://dl.acm.org/doi/10.1145/1557019.1557077

[22] C. Suen, S. Huang, C. Eksombatchai, R. Sosic, and J. Leskovec, "NIFTY: A system for large scale information flow tracking and clustering," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, pp. 1237–1248. [Online]. Available: https://dl.acm.org/doi/10.1145/2488388.2488496

[23] G. Colavizza, M. Infelise, and F. Kaplan, "Mapping the Early Modern News Flow: An Enquiry by Robust Text Reuse Detection," in *Social Informatics*, ser. Lecture Notes in Computer Science, L. M. Aiello and D. McFarland, Eds. Springer International Publishing, pp. 244–253.

[24] S. Vakulenko, M. Gobel, A. Scharl, and L. Nixon, "Visualising the Propagation of News on the Web."

[25] G. A. Miller, "WordNet: A lexical database for English," vol. 38, no. 11, pp. 39–41. [Online]. Available: https://dl.acm.org/doi/10.1145/219717.219748

[26] D. P. A. Corney, D. Albakour, M. Martinez-Alvarez, and S. Moussa, "What do a Million News Articles Look like?" [Online]. Available: https://ceur-ws.org/Vol-1568/paper8.pdf

[27] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," pp. 45–50.

[28] Q. V. Le and T. Mikolov. Distributed Representations of Sentences and Documents. [Online]. Available: http://arxiv.org/abs/1405.4053

[29] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. [Online]. Available: http://arxiv.org/abs/1310.4546

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. [Online]. Available: http://arxiv.org/abs/1301.3781

[31] J. H. Lau and T. Baldwin. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. [Online]. Available: http://arxiv.org/abs/1607.05368

[32] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo, "*SEM 2013 shared task: Semantic Textual Similarity," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Association for Computational Linguistics, pp. 32–43. [Online]. Available: https://aclanthology.org/S13-1004

[33] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, "SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity," in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Association for Computational Linguistics, pp. 385–393. [Online]. Available: https://aclanthology.org/S12-1051

[34] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, R. Mihalcea, G. Rigau, and J. Wiebe, "SemEval-2014 Task 10: Multilingual Semantic Textual Similarity," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, pp. 81–91. [Online]. Available: https://aclanthology.org/S14-2010

[35] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, G. Rigau, L. Uria, and J. Wiebe, "SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, pp. 252–263. [Online]. Available: https://aclanthology.org/S15-2045

[36] E. Agirre, A. Gonzalez-Agirre, I. Lopez-Gazpio, M. Maritxalar, G. Rigau, and L. Uria, "SemEval-2016 Task 2: Interpretable Semantic Textual Similarity," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, pp. 512–524. [Online]. Available: https://aclanthology.org/S16-1082

[37] A. Huang, "Similarity measures for text document clustering."

[38] N. Bhatia and Vandana. Survey of Nearest Neighbor Techniques. [Online]. Available: http://arxiv.org/abs/1007.0085

[39] P. Berkhin, "A Survey of Clustering Data Mining Techniques," in *Grouping Multidimensional Data: Recent Advances in Clustering*, J. Kogan, C. Nicholas, and M. Teboulle, Eds. Springer, pp. 25–71. [Online]. Available: https://doi.org/10.1007/3-540-28349-8_2

[40] O. Hassanzadeh, P. Awasthy, K. Barker, O. Bhardwaj, D. Bhattacharjya, M. Feblowitz, L. Martie, J. Ni, K. Srinivas, and L. Yip, "Knowledge-Based News Event Analysis and Forecasting Toolkit," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, pp. 5904–5907. [Online]. Available: https://www.ijcai.org/proceedings/2022/850

[41] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik, "Event registry: Learning about world events from news," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14 Companion. Association for Computing Machinery, pp. 107–110. [Online]. Available: https://dl.acm.org/doi/10.1145/2567948.2577024

[42] B. D. Trisedya, G. Weikum, J. Qi, and R. Zhang, "Neural Relation Extraction for Knowledge Base Enrichment," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 229–240. [Online]. Available: https://aclanthology.org/P19-1023

[43] B. D. Trisedya, J. Qi, and R. Zhang, "Entity Alignment between Knowledge Graphs Using Attribute Embeddings," vol. 33, no. 01, pp. 297–304. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/3798

[44] P. Mell and T. Grance, "The NIST Definition of Cloud Computing."

[45] Australian Digital Observatory -. [Online]. Available: https://www.ado.eresearch.unimelb.edu.au/

[46] L. Morandini, A. R. Mohammad, and R. O. Sinnott, "MAPPING THE CHATTER: SPATIAL METAPHORS FOR DYNAMIC TOPIC MODELLING OF SOCIAL MEDIA," vol. XLVIII-4/W1-2022, pp. 315–320. [Online]. Available: https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLVIII-4-W1-2022/315/2022/

[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [Online]. Available: http://arxiv.org/abs/1810.04805

[48] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books," pp. 19–27. [Online]. Available: https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Zhu_Aligning_Books_and_ICCV_2015_paper.html

[49] M. V. Koroteev. BERT: A Review of Applications in Natural Language Processing and Understanding. [Online]. Available: http://arxiv.org/abs/2103.11943

[50] D. Q. Nguyen, T. Vu, and A. T. Nguyen. BERTweet: A pre-trained language model for English Tweets. [Online]. Available: http://arxiv.org/abs/2005.10200

[51] M. Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. [Online]. Available: http://arxiv.org/abs/2203.05794

[52] Melbourne Research Cloud Documentation. [Online]. Available: https://docs.cloud.unimelb.edu.au/guides/access_from_nectar/

[53] ARDC Nectar Research Cloud — ARDC. https://ardc.edu.au/. [Online]. Available: https://ardc.edu.au/services/ardc-nectar-research-cloud/

[54] K. Makice, *Twitter API: Up and Running: Learn How to Build Applications with the Twitter API*. "O'Reilly Media, Inc.".

[55] J. C. Anderson, J. Lehnardt, and N. Slater, *CouchDB: The Definitive Guide: Time to Relax*. "O'Reilly Media, Inc.".

[56] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," vol. 51, no. 1, pp. 107–113. [Online]. Available: https://dl.acm.org/doi/10.1145/1327452.1327492