Automated Concern Exploration in Pandemic Situations - COVID-19 as a Use Case

No Author Given

No Institute Given

Abstract. The recent outbreak of the coronavirus disease (Covid-19) rapidly spreads across most of the countries, which has significantly impacted the people's life in almost every aspect and challenged an economy strain around the world. To alleviate the panics and prevent any potential social crisis, it is essential to effectively detect the public concerns through the social media. Twitter, a popular online social network, allows people to share their thoughts, views and opinions towards the latest events or news. In this study, we proposed a deep learning-based framework to explore public concerns for Covid-19 automatically, where Twitter has been utilised as the key source of information. We extracted and analysed public opinions towards the pandemic. Furthermore, as part of the proposed framework, a knowledge graph of public concern has been constructed to investigate the interconnections.

Keywords: Covid-19 · Concern exploration · Twitter · Opinion mining · Knowledge graph

1 Introduction

The outbreak of Covid-19 has been recognised as a severe global threat by the World Health Organisation (WHO), and it has been impacting the world in various aspects [21]. People in some infected countries have been advised to stay at home, avoiding public and crowded places; large convocations have been cancelled; cities have been placed under lock-down and schools have been shut down [7]. A striking particularity of this pandemic is the spread of both physical disease and mental panic, i.e., not only did the coronavirus itself spread very rapidly but so did the related information about the outbreak [5, 11]. People tend to express their opinions and concerns about this crisis through one of the most prevalent communicational channels, i.e., online social networks [19, 12]. Such information provides strong evidence on what public concerns to be addressed urgently. This also explicitly reveals where to intervene with key communication campaigns, what guidelines for protection to be provided and how to determine the measure for alleviating public fear [17, 9, 4]. Therefore, it is essential to automatically capture and detect the public concerns in this pandemic situation.

Concern reveals negative beliefs about worry. Thus, it is essential to extract people's opinions, sentiments, emotions from their generated texts for concern detection. Opinion mining, a sub-field of Natural Language Processing (NLP),

has been recognised as an important role in making sense of the textual data, e.g., Tweets from Twitter [1] [22]. In particular, sentiment analysis and opinion mining has been widely adopted in marketing, assisting customers in making purchasing decisions and aiding the business owners to understand the customers' feedback [22, 20]. Likewise, in the context of Covid-19, it turns out to be an essential technique to analyse the public opinions and detect corresponding concerns. On the other side, the interconnected concerns exploration is also useful. By considering the causality, formation of a concern potentially results in other concerns. Because a group of people's concern may lead a chain of actions, that damage the interests of another group people. For example, the poor concern about the living expenses, while the rich get worried about being robbed by the poor. In this sense, mitigating one concern can potentially address the related ones. Knowledge Graph (KG) has been recognised as a suitable technique to collate all the facts and create interconnected results. Therefore, KG has been employed in our approach as an essential components unravel the associations among public concerns.

In this paper, we formally defined public concern and proposed an automated approach to detect concerns by facilitating deep-learning models.

2 Related Works

Due to the severity and rapidly-evolving nature of the Covid-19, the response from both computer science and data science community has been ongoing. A number of researchers collected and published Covid-19 datasets, which greatly contributes to the studies of online conversation dynamics in the context the epidemic outbreak [1, 3, 15, 16, 24]. Based on these open datasets, further research works around content analysis have been conducted. Boberg et al. analyse the factual basis of public fears based on the alternative news media's output on Facebook [2] during the early Corona crisis [2]. Jahanbin et al. analyse corpus of Covid-19 tweets and identify common responses to the pandemic over time [13]. Li et al. attempt identify situational information of Covid-19 and explore it is being disseminated on social media [18]. Most of researchers conduct statical analysis in regards to the keywords and trending topics. However, very few research works have been dedicated to mining the fine-grained opinions for exploring and monitoring the social concerns. Different from the existing studies, we develop a novel model, which is capable of analysing the aspect-based opinions and detecting the public concerns through social media.

Deep learning has demonstrated great success on solving complex learning problems. Nowadays, it has been widely applied to NLP and social media data filtering, which yields remarkable performance [14]. Jelodar et al. utilise NLP and Recurrent Neural Network (RNN) for Covid-19 related discussion extraction and topic modelling from social media, where a supervised method is developed to

[1] https://twitter.com/
[2] https://www.facebook.com/

analyse people opinion of Covid-19-related comments from reddit forums [3] [14]. Fan et al. propose a hybrid machine-learning pipeline for automated mapping of events and locations from social media posts during disasters, where a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) has been adopted to classify the posts [8]. However, very few researchers develop deep learning models for social concern detection in the context of Covid-19. Moreover, the connections of public concerns are also missing from the existing studies.

Knowledge Graph (KG) has been acknowledged as a significant technique to collate all the facts and present interconnected results. It has been adopted in many AI and NLP tasks, e.g., Q/A, machine translation, topic extraction, etc. Gottschalk et at. studied a multilingual event-centric temporal KG, i.e., EventKG, aiming at addressing insufficient coverage and completeness with respect to events and temporal relations of popular knowledge base database including DBpedia, Yago and Wikidata [10]. Whereas, the latest news and events cannot be covered since the proposed approach relies on the existing static knowledge base, thus may not cope with a dynamic information source. Yoo et al. develop a novel method, i.e., PolarisX, to automatically expand a KG by utilising a fine-tuned pre-trained multilingual BERT model, where the data are constantly collected from news sites and social media platforms in real-time [23]. Based on these studies, we employ KG as an essential components to explore the associations among public concerns in real-time.

To the best of our knowledge, this is the first research study, applying NLP and deep learning algorithms to explore public concerns and deeply investigate the corresponding relations using online social media data in the context of Covid-19.

3 Preliminaries

In this section, we formally define public concerns revealed from online social network, i.e., Twitter, as well as the concern detection problem.

3.1 Formal Definition

Definition 1: Entity generally describes an object which can be referred to as a product, medical service, drug, disease, person, event or organisation. In the current setting, an entity represents a word object which is associated with the public concern. Specifically, any tweet $tw_i, tw_i \in TW$ posted by user consists of a word sequence $tw_i = \{w_1, w_2, ..., w_n\}$, where TW denotes the collected tweet data. An entity $e_i \in E$ appears to be an element from the word sequence, semantically relating to the public concerns, where $E = \{e_1, e_2, ..., e_n\}, n \in \mathbb{N}$ denotes the entity set of the corpus.

Entity and tweet represent a many-to-many relationship, i.e., an entity may exist in many tweets, and a tweet may include more than one entities.

[3] https://www.reddit.com/

Definition 2: Concern in general refers to negative beliefs about worry. In this study, concern is defined as most people's negative opinions towards a particular aspect of a crisis. Mathematically, a concern c_i identified from tw_i is denoted by a quadruple $c_i = (e_j, r_j, p_j, t_j), c_i \in C$, where e_j represents the entity extracted from tweet tw_i, r_j means the re-tweet number, $p_j \in \{Positive, Negative, Neutral\}$ is the sentiment polarity towards tw_i, and t_j shows the posting date and time of the tweet.

Definition 3: Concern Triple is recognised as the fundamental element, contributing to the formation of a public concern knowledge graph. A single word or phrase turns out to be very limited to present the true meaning of a concern. By contrast, the triple structure is capable of semantically revealing what a concern is about. A public concern $ct_i = (s_i, r_i, o_i)$ comes with three components s_i, r_i and o_i, referring to as subject, relationship and object, respectively.

For example, *"The Corona Virus is like an intelligence test for humanity"* is a tweet posted by one user at 21:59:30 19th March 2020 and the re-tweet number is 385. Based on these facts, the identified entities $e1$=*corona virus*, $e2$= *intelligence test*, r_1=r_2=385, p_1=p_2=0.01, t_1=t_2=*21:59:30 19th March 2020*, and identified concerns $c_1 = (e_1, r_1, p_1, t_1)$ and $c_2 = (e_2, r_2, p_2, t_2)$. The corresponding concern triple is:

$$ct_1 = (corona\ virus, is\ like, intelligence\ test)$$

3.2 Problem Formulation

Based on the definitions given in the previous section, we formally define the public concern detection. It aims identify a set of typical concerns $\{c_m | c_m \in C \wedge p_m = Negative\}$ that attract most people's attention from a large amount of social media data TW, explain the meanings by extracting semantic concern triples and exploring the interconnections among detected concerns $\{(s_m, r_m, o_m) | s_m, o_m \in C \wedge \exists r_m\}$.

4 Automated Public Concern Detection

In this section, we systematically elaborate our proposed automated concern detection framework. The explanations are developed mainly based on the social media data flow.

Figure 1 demonstrates the proposed automated concern detection framework, which turns out to be a hierarchical multi-stage processes with four main components, including data pre-processing, deep learning models, typical concern extraction and clustering, and concern knowledge graph construction.

4.1 Data Pre-processing and Information Extraction

Twitter has been reconsigned a popular public online social platform for people to communicate and share opinions with each other. However, due to informal and abbreviated words, as well as improper sentences with grammatical errors, it is almost impossible to analyse tweets or extract meaningful information without adequate data pre-processing.

The initial module of the framework tends to improve the quality the social media data through pre-processing. The typical actions incorporate removing meaningless data fractions, such as URL, hash-tag, punctuation, non-English contents, symbols, stop words and emoji.

As for the information extraction, three elements, i.e., concern entities, sentiment score and re-tweet number, are supposed to be captured. The framework detect concern entities through deep learning models. Sentiment score also plays an important role in determining whether the a twitter statement is concerned. The concern sentiment score in the current context can be obtained by combining text sentiment score with the emoji sentiment score based on Emoji Sentiment Ranking [4]. Re-tweet count appears a significant indicator, showing the popularity of a tweet and reflecting the degree of public attention.

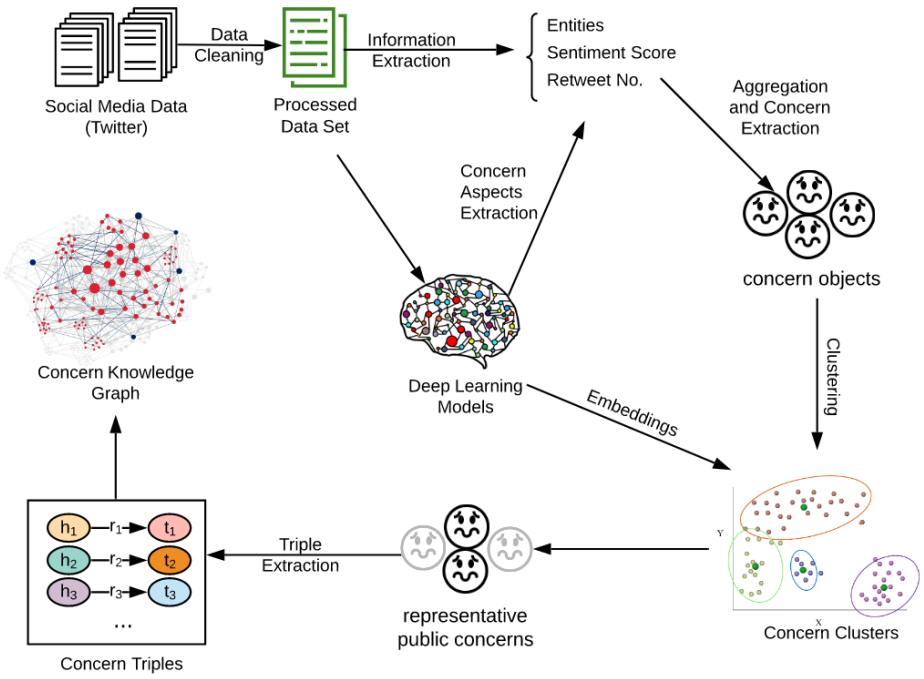

Fig. 1. Automated Concern Detection Framework

[4] Emoji Sentiment Ranking `http://kt.ijs.si/data/Emoji_sentiment_ranking/`

4.2 Deep Learning Models

Deep learning models have been utilised as a semantic analyser in the automated concern detection framework. BERT appears to be a pre-trained language model released by Google, which is capable of taking up a number of classic NLP tasks [6, 23]. In the current setting, we fine-tune the pre-trained BERT model for two purposes, i.e., extracting potential public concern entities and create embeddings for the detected concerns.

In the former, to train a concern entities extraction model, it is required to figure out the potential concerns in a tweet. Thus, the annotated concern entities and tweets are used to train the model.

The input representation of BERT for each token is summation of the token, segment and position embeddings. $[CLS]$ is appended to the beginning of each sentence and $[SEP]$ is inserted in the sentence end to separate two sentences.

For a tweet tw with entity e, we add $[CLS]$ and $[SEP]$ to the beginning and ending of tw, and a special token '$' before and after entity to make BERT module capture the location information.

For example, the original tweet is *"The Corona Virus is like an intelligence test for humanity"*, after inserting special separate tokens, it will become *"[CLS] The $ Corona Virus $ is like an $ intelligence test $ for humanity [SEP]"*.

There are total 12 encoder layers for BERT module and we use summation of last 4 layer output as entity embedding as equation (2). So the final embedding for each entity is as below.

$$H_i = \frac{1}{(m-n+1)*4} \sum_{k=n}^{m} \sum_{l=L8}^{L11} h_k^{(l)} \tag{1}$$

where H_i is the embedding of i^{th} entity, and $h_k^{(l)}$ refers to the embedding of $k^{(th)}$ token at layer k.

4.3 Concern Extraction and Clustering

K-means clustering is an unsupervised learning algorithm which is to find groups in the data with given cluster number K.

$$\underset{o_k \in O}{argmin} dist(o_k, H^{(i)})^2, \tag{2}$$

where o_k is the k^{th} cluster centroids in set O with $k \in K$. $dist(.)$ is the standard(L_2) Euclidean distance. $H^{(i)}$ is the sum of last 4 layer output of BERT for i^{th} concern entity.

$$H^{(i)} = \frac{1}{4} \sum_{i} h^{(i)} \tag{3}$$

Where $h^{(i)}$ is the BERT i^{th} layer output and $i = \{8, 9, 10, 11\}$.

For all extracted concerns in previous step, the entity embeddings are as input of K means clustering algorithm. In order to determine the optimal value of k, Silhousette analysis is used to study the separation distance between resulting concern clusters.

$$SilhouetteCoefficient = \frac{(b_i - a_i)}{max(b_i, a_i)} \tag{4}$$

where b_i is the mean intra cluster distance to other points in the same cluster and a_i depicts mean distance to other points to its closest neighbouring cluster. The Silhousette score displays a measure of how close each point in one cluster is to points in the neighbouring clusters, and its value range is [-1, 1] in which +1 means points in one cluster are far away from the neighbouring clusters.

4.4 Concern Knowledge Graph

After concern clustering, public representative concerns are selected as candidate concerns for concern triple extraction using syntactic dependency parser of Spacy. The concern triplet subject-predicate-object is denoted as $(head, relationship, tail)$. We will detect relationship, head or tail based on extracted concern, which means if concern is subject in tweet, the task is to determine object and relation in same tweet.

For example, in tweet "$PrinceCharleshastestedpositiveforCorona$." the extracted concern is "$PrinceCharles$" as head of triple, then determine predicate "$tested$" as relationship and object "$positiveCorona$" as tail.

In order to easily understand the meaning about extracted concerns, concern knowledge graph is constructed using concern triples.

4.5 NLP - Deep Learning Module

@Jingli, needs the explanations with formulas

4.6 KG - Construction Module

5 Experiments

@Jingli, IN THIS SECTION, PLEASE EVALUATE THE MODEL, THE PERFORMANCE OF ASPECT EXTRACTION AND OPINION MINING. IT WOULD BETTER IF SIMPLE COMPARISON AGAINST THE OTHER CLASSIC ALGORITHMS COULD BE GIVEN.

5.1 Dataset

Dataset [5] [16]

[5] https://ieee-dataport.org/open-access/corona-virus-covid-19-tweets-dataset

Rabinadra collect COVID-19 related English tweets from Twitter by using keyword filter, like "corona", "coronavirus", "covid", "covid19" and variants of "sarscov2". We use software Hydrator to download full tweet information including columns "created at", "retweet count","text","user location" according to provided Tweet IDs.

5.2 Evaluation

5.3 Covid-19 Analysis

6 Conclusion and Future Work

combination of NLP and KG Future work can be: disaster relief, fake news detection, disaster management and mitigation

References

1. Covid-19 open research dataset (cord-19). `https://pages.semanticscholar.org/coronavirus-research` (2020, March 20), [Online; accessed 11-April-2020]
2. Boberg, S., Quandt, T., Schatto-Eckrodt, T., Frischlich, L.: Pandemic populism: Facebook pages of alternative news media and the corona crisis–a computational content analysis. arXiv preprint arXiv:2004.02566 (2020)
3. Chen, E., Lerman, K., Ferrara, E.: Covid-19: The first public coronavirus twitter dataset. arXiv preprint arXiv:2003.07372 (2020)
4. Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C.M., Brugnoli, E., Schmidt, A.L., Zola, P., Zollo, F., Scala, A.: The covid-19 social media infodemic. arXiv preprint arXiv:2003.05004 (2020)
5. Depoux, A., Martin, S., Karafillakis, E., Preet, R., Wilder-Smith, A., Larson, H.: The pandemic of social media panic travels faster than the covid-19 outbreak. Journal of Travel Medicine (2020)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Enberg, J.: Covid-19 concerns may boost ecommerce as consumers avoid stores. `https://www.emarketer.com/content/coronavirus-covid19-boost-ecommerce-stores-amazon-retail` (2020, March 10), [Online; accessed 11-April-2020]
8. Fan, C., Wu, F., Mostafavi, A.: A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters. IEEE Access (2020)
9. Garrett, L.: Covid-19: the medium is the message. The Lancet **395**(10228), 942–943 (2020)
10. Gottschalk, S., Demidova, E.: Eventkg: A multilingual event-centric temporal knowledge graph. In: European Semantic Web Conference. pp. 272–287. Springer (2018)
11. Hou, Z., Du, F., Jiang, H., Zhou, X., Lin, L.: Assessment of public attention, risk perception, emotional and behavioural responses to the covid-19 outbreak: social media surveillance in china. Risk Perception, Emotional and Behavioural Responses to the COVID-19 Outbreak: Social Media Surveillance in China (3/6/2020) (2020)

12. Hu, Y., Bai, Q., Li, W.: Context-aware influence diffusion in online social networks. In: Pacific Rim Knowledge Acquisition Workshop. pp. 153–162. Springer (2019)
13. Jahanbin, K., Rahmanian, V.: Using twitter and web news mining to predict covid-19 outbreak (2020)
14. Jelodar, H., Wang, Y., Orji, R., Huang, H.: Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. arXiv preprint arXiv:2004.11695 (2020)
15. Killeen, B.D., Wu, J.Y., Shah, K., Zapaishchykova, A., Nikutta, P., Tamhane, A., Chakraborty, S., Wei, J., Gao, T., Thies, M., et al.: A county-level dataset for informing the united states' response to covid-19. arXiv preprint arXiv:2004.00756 (2020)
16. Lamsal, R.: Corona virus (covid-19) tweets dataset (2020). https://doi.org/10.21227/781w-ef42, http://dx.doi.org/10.21227/781w-ef42
17. Lazard, A.J., Scheinfeld, E., Bernhardt, J.M., Wilcox, G.B., Suran, M.: Detecting themes of public concern: a text mining analysis of the centers for disease control and prevention's ebola live twitter chat. American journal of infection control **43**(10), 1109–1111 (2015)
18. Li, L., Zhang, Q., Wang, X., Zhang, J., Wang, T., Gao, T.L., Duan, W., Tsoi, K.K.f., Wang, F.Y.: Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo. IEEE Transactions on Computational Social Systems (2020)
19. Li, W., Bai, Q., Zhang, M., Nguyen, T.D.: Automated influence maintenance in social networks: an agent-based approach. IEEE Transactions on Knowledge and Data Engineering **31**(10), 1884–1897 (2018)
20. Rathan, M., Hulipalled, V.R., Venugopal, K., Patnaik, L.: Consumer insight mining: aspect based twitter opinion mining of mobile phone reviews. Applied Soft Computing **68**, 765–773 (2018)
21. Sohrabi, C., Alsafi, Z., ONeill, N., Khan, M., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, R.: World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19). International Journal of Surgery (2020)
22. Wang, R., Zhou, D., Jiang, M., Si, J., Yang, Y.: A survey on opinion mining: from stance to product aspect. IEEE Access **7**, 41101–41124 (2019)
23. Yoo, S., Jeong, O.: Automating the expansion of a knowledge graph. Expert Systems with Applications **141**, 112965 (2020)
24. Yu, J.: Open access institutional and news media tweet dataset for covid-19 social science research. arXiv preprint arXiv:2004.01791 (2020)