News Event Prediction using Causality Approach on South China Sea Conflict

Teo Wen Long
Cyber Threat Intelligence Lab, Information Assurance & Security Research Group (IASRG), School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia
81310 Johor Bahru, Johor, Malaysia
teowenlong0316@gmail.com

Anazida Zainal
Cyber Threat Intelligence Lab, Information Assurance & Security Research Group (IASRG), School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia
81310 Johor Bahru, Johor, Malaysia
anazida@utm.my

Mohd Nizam Kassim
Cyber Security Responsive Service Division, CyberSecurity Malaysia, Level 7 Tower 1, Menara Cyber Axis, Jalan Impact, 63000 Cyberjaya,
Selangor, Malaysia
nizam@cybersecurity.my

2021 3rd International Cyber Resilience Conference (CRC) | 978-1-6654-1844-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/CRC50527.2021.9392431

***Abstract*—South China Sea (SCS) generates huge economic value in fishing and shipping lane as well as a high amount of natural resources. Due to its strategic location and high revenue generated, SCS became a place where several nearby countries competed for its territorial claims. Famous territorial disputes such as Spratly islands, Paracel island, Scarborough Shoal happened due to claim on SCS wealth. Newspapers are the main medium that disseminate the message to the public and update whenever SCS conflict happens. News related to SCS events or conflicts usually contain causal relationships between cause and effect. This causal relationship can be extracted and analyzed to obtain the trends of events and conflicts that have happened. In order to avoid any inevitable conflict among countries in SCS region, event prediction is important as it gives a better insight and foresee future events that might happen. In this paper, phrase similarity is used as important metrics for prediction models. First, it extracts news articles based on causality connectors such as "because", "after", "lead to", etc. into <cause, effect> tuple. Then, three different embedding techniques, Doc2vec, InferSent and BERT were evaluated based on their best similarity score. The selected embedding technique is used to construct the prediction model and predict South China Sea conflict related events. A crude prediction is done based on similarity of past causes. The result shows that BERT has the highest average accuracy of 50.85% in getting the most similar phrase. By using the causal prediction model, a future possible event can be predicted and this helps to increase the awareness of national security among SCS nearby countries.**

Keywords*—South China Sea Conflict, Event Prediction, Causality, Sentence Embedding, Sentence Similarity

I. INTRODUCTION

South China Sea (SCS) is a conflict zone whereby an estimated USD5 trillion worth of raw products shipped through shipping lanes in SCS each year and its nearby countries made them fight over each other to have the main control of the whole SCS. [1]. The conflict is known as South China Sea disputes. The events of territorial disputes of South China Sea are reported in mainstream newspapers and tabloids as well. For reliable news, researchers opt to use articles published by national news agencies. SCS conflicts raises concerns about the onset of world war as for example, a near-collision between US warship, Decatur and Chinese Luoyang missile destroyer in South China Sea highlights the escalating danger of confrontation between US and China. [2]. Hence, every country has its obligation to protect its national security.

Event prediction is a data analytic technique that makes use of experience and knowledge as well as patterns from the past to predict future events. Meanwhile, causality is the relationship between cause and effects. Every event that occurs, has a pairing pattern, cause and followed by its effect. In order to have an foresight about these disputes, event prediction is necessary, and causality should be taken as main attributes. However, there are still several problems and challenges need to be solved in order to achieve a good prediction model based on causality.

This paper aims to study the performance between sentence embedding techniques and develop an accurate causal prediction model to obtain the most possible causal event. This paper is organized into six sections. Section one covers introduction and gives a background to the problem followed by Section two, which discusses the literature review. Meanwhile Section three describes the methodology used in this study and Section four illustrates our design and implementation. Finally, Section four discusses the results obtain and Section five concludes the paper.

A. Problem Background

National security is always the top priority of governments to protect sovereignty of their countries. There are many aspects of national security such as territorial, economic, physical, social, political etc. Due to geological and resources advantages of the South China Sea, countries within the region such as Brunei, China, Taiwan, Malaysia, Indonesia, Philippines, Vietnam etc. made competing territorial claims over it. Based on news on The National Interest in 2016, an estimated of US 5 trillion worth of global trade passes through the South China Sea annually. Hence, territorial disputes in the South China Sea started to concern the worldwide community. In order to claim the ownership of the South China Sea, countries are challenging each other by putting military force in the area. This can be observed from the news of China spending almost 1 year to build 7 new islands by moving sediment from the sea floor to reefs and after that

focused on building ports, airstrips and other military structures on the islands.

South China Sea dispute had a brief background involving a timeline from 221 BC until recently. Each of the historical events occurs and accumulates and eventually things go haywire. Many dispute events happened on either small or large scales. For example, Spratly island dispute [3] and the "nine-dash" line [4] that was initialized by China are some of the significant disputes in the South China Sea. Besides these two issues, there are many issues that remain unsolved and will constantly concern the worldwide community.

However, all the information retrieved from news articles are unstructured. Unstructured data have no recognizable structure via pre-defined data models and schema and are mainly generated by humans or machines. [5] By collecting these unstructured data from the past and analyzing its trends, we are able to have a better understanding about what may happen in the future. In SCS disputes, event prediction is important to give the public a better understanding about future events that might happen. A better policy can be made with regard to protecting national security under SCS disputes with the event prediction technique based on unstructured data in news articles.

In event prediction based on news articles, there are some challenges. First, a news article is unstructured data that contains a lot of valuable information in terms of cultural, social and historical [6] and not suitable to be stored into traditional row and column structure of relational databases. It requires substantial manual effort to analyze and extract the essential information from news articles. Second, an event that causes another event may be completely different from the real prediction. It is indicating that the predictive model provides a inaccurate outcome.

There are several researchers working on topic event prediction with different methods. [20] proposed a predictive neural network model that learns embeddings for words describing events, a function to change embeddings into event representation and a function to predict the degree of relationship between two events. However, the model is more focused on chain or events sequence which is good for rich-informative events but might not be suitable for news articles that have unordered sequence. [21] proposed an event prediction model for Tweets using temporal sentiment analysis and causal rules extraction. This model is useful to analyze a user's sentiments and predict future events using temporal attributes. This study analyzes sentiment of user's opinion and is not suitable for news articles whereby formal news reports seldom express their sentiments within the articles.

The current research done is more focused on casual event detection and extraction, which is related to effective distributed word or sentence representation. [7] had proposed a state-of-art framework, Word2Vec for distributed word representation. However, Word2Vec is limited to those words that are morphologically similar where Word2Vec embed every word as an independent vector.

Thus, this research is aimed to provide a prediction model that attempts address these problems by using state-of-art sentence representation and sentence similarity. In this study, we trained and compared three different sentence representation techniques (Doc2Vec, InferSent, and BERT) as well as their sentence similarity with the input cause phrase in order to get the most possible predicted output and develop a causal prediction model based on best-fit embedding technique.

B. Problem Statement

South China Sea (SCS) is a conflict zone where events happen from time to time with different severity. This greatly impacts or influences the policy made by the government of Malaysia to overcome the negative impacts brought by SCS territorial disputes in terms of national security. The problem is to extract valuable information from SCS conflict events and predict the future events that may happen. Besides, online news is unstructured data and extracting correct information from massive online news articles that contain different resources automatically is part of the challenges.

II. Literature REview

A. Event Prediction

Event prediction is a technique to measure the trend of happening events and forecast upcoming events that might happen. Events prediction is relevant in many domains and researches in this area are active and wide. Examples of domains in event prediction are a) Natural disaster such as earthquake [8] and tsunami [9], b) Political such as election prediction [10], c) Economics such as predicting market stock price [11]. News articles, report many events that happen almost every day. From political issues, to social problems, economic trends and entertainments, the newspaper provides public daily updates on these events. Recently, conflict on the South China Sea (SCS) has become popular and continuously escalated. The events related to SCS conflict concerns the surrounding countries and actions should be taken to prevent any serious tragedy that impact the sovereignty of these countries.

B. South China Sea Conflict

South China Sea (SCS) is located strategically within Asia countries which are; China, Taiwan, Indonesia, Philippines, Vietnam, and Malaysia. It has the busiest shipping laneas one-third of the world's shipping passes through SCS and almost 3.37 trillion global trade happened within SCS in 2016 [12]. Due to this huge market value, many countries started to claim that SCS is their own territory and China even illustrated a "nine-dash line" which is a huge part of SCS and claim that region within the "nine-dash line" is China's territory [13]. For every dispute, the public are able to know the flow of events through newspapers since these are published publicly. Hence, there are huge amounts of valuable information that represent and reflect the actual events in SCS disputes.

C. Event Detection and Extraction

Annotation is an important concept in obtaining useful information in Natural Language Processing (NLP). By annotating raw text data, all the text becomes meaningful as a tag with specific characteristics. For example, a Named Entity Recognition (NER) annotation can simply help researchers to eliminate most of the noise in the raw data by only obtaining Person, Location, Time, etc. The most widely used annotation tools for NLP are Brat and WebAnno. Figure 2 shows the screenshot of Brat annotation tools. By using Brat, cause and effect can be easily annotated followed by their relationship (cause-effect). In this study, Brat annotation tools was used to extract the cause and effect from the selected sentences and create <cause, effect> causality pairs.

Causal cue word is one of the good ways to detect event causality from sentences. Sentences can be split into candidate causal and effect phrases using a set of causal cue words. Figure 1 shows the example of causal cue words. In this study, causal cue words are used to obtain sentences with event causality for further processing.

cause	because	lead to	leads to
result	due to	make	since
so	increase	therefore	after

Figure 1 Causal Cue Words [14]

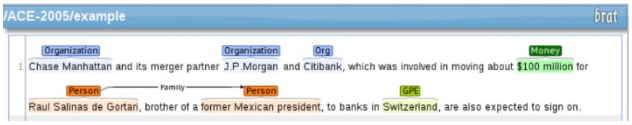

Fig. 2 Screenshot of Brat Annotation Tool [22]

D. Event *Representation*

There are 3 state-of-art sentence representation techniques studied in this study, Doc2Vec, InferSent and BERT. Each of the techniques contains unique characteristics and benefits for different use cases.

1) *Doc2Vec*

Doc2Vec is proposed by Le and Mikolov [15]. It is also known as Paragraph Vector as the algorithm adds a Paragraph ID on its training. Based on the Word2Vec approach, Doc2Vec follows the concept of CBOW and skip-gram, additionally added another vector (Paragraph ID). There are also 2 methods that are similar to CBOW and skip-gram in Word2Vec which are PV-DM (Paragraph Vector - Distributed Memory) and PV-DBOW (Paragraph Vector - Bag-of-Words). Figure 3 shows the illustration of PV-DM and PV-DBOW models. PV-DM predicts the missing target word in the context while PV-DBOW predict all the context from the target word.

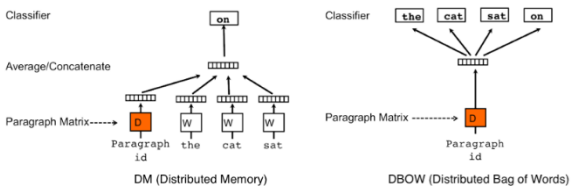

Fig. 3 PV-DM and PV-DBOW model in Doc2Vec [15]

2) *InferSent*

InferSent is a one of the state-of-art sentence embedding technique that provides semantic sentence representation. Facebook researchers had proposed InterSent in [16]. InferSent first embeds the sentence using sentence encoder. There are several techniques in the sentence encoder and one of the most effective techniques is Bi-directional LSTM network with max or mean pooling. Each vector is concatenated between forward LSTM and a backward LSTM that is able to read sentences in the opposite direction. Then, max or mean pooling is used in these concatenated vectors to form fixed-length vectors. InferSent makes use of Stanford Natural Language Inference (SNLI) dataset and manually label with 3 categories (entailment, contradiction and neural), then create NLI classifier. NLI classifier is a method of finding directional relationships between text fragments and will then be used to extract the relations between text and hypothesis.

3) *BERT*

BERT (as shown in Figure 4) stands for Bidirectional Encoder Representation from Transformers and it is a state-of-the-art sentence embedding technique proposed by Google AI language. [17]. BERT makes use of Transformer, an attention mechanism that learns contextual relation between words in a text. Figure 4 shows the BERT input representation. There are 2 important elements in BERT, Masked LM (MLM) and Next Sentence Prediction (NSP). MLM allows bidirectional training where the model uses the context words surrounding a [MASK] token and tries to predict what the [MASK] word should be. Besides, in NSP, [CLS] token is inserted at the beginning of the first sentence while [SEP] token is inserted at the end of each sentence. Combining with [MASK] token, it forms a sequence and the entire sequence is embedded through the Transformer model.

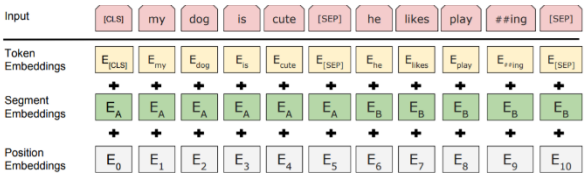

Fig. 4 BERT input representation [17]

E. *Vector Similarity*

Vector similarity is one of the important measures for word vector or sentence vector. Example of technique measuring vector similarity are Cosine Similarity, Euclidean distance, Jaccard distance and word mover's distance. Among these techniques, Cosine similarity is being widely used and applied in NLP domain. [18] [19]. Cosine similarity is used for chatbot engine. When user input a random query, the input is embedded into vector and compute cosine similarity calculation with every vector in the model. After that, vector with highest similarity is selected and returning the original question. The question is said to have the most similar textual semantic to the input query. This question is then linked user to the corresponding answer. The formula for cosine similarity is shown in Eq. (1), where A and B are 2 vector attributes.

$$\text{Similarity} = \cos(\theta) = \frac{A.B}{||A|| * ||B||} \quad (1)$$

III. METHODOLOGY AND IMPLEMENTATION

Figure 5 shows the research framework of this study. There are 3 phases which are a) Development of causality data corpus using South China Sea Conflict news article, b) Phrase representation and Phrase Similarity Measure and c) Causality Prediction Model.

This study focuses on South China Sea conflict news articles from online resources, China and Vietnam. Keywords such as "China", "Vietnam", "South China Sea", "Spratly", "World Court", etc. are used to search online news articles. Once the news article matches with the keywords, it is crawled as our data sets for further data processing. The collected data is then grouped according to the corresponding country based on the location of the news published. A total of 200 article are

crawled and cleaned to raw text. This text is separated into sentence level and annotated into <Cause, Effect> causality phrase pairs. Figure 6 shows the example of news articles crawled from online news. All the data is separated into ratios of 80% and 20% to training data and testing data respectively. Besides, a benchmark dataset, SemEval 2010 Task 8 is used for validation. In SemEval 2010 Task 8, all the CauseEffect entities are extracted and there are 557 training data and 140 testing data.

To measure the performance of the prediction model, predicted truth and accuracy are used as main evaluation indicator. Predicted truth is defined as binary measurement for correctly predicted events. For each predicted event in a causal pair, if it is correctly predicted, the predicted causal pair is labelled as 1 and 0 otherwise. All of these labels are treated as predicted truths of the model. Accuracy of the model is also used as an evaluation metric. Accuracy of the model indicates the suitability of the embedding technique. Higher accuracy means it is more suitable. The formula for accuracy measure of the model is as follows:

$$\frac{Total\ number\ of\ correctly\ predicted\ effects}{Total\ number\ of\ effects} \quad (2)$$

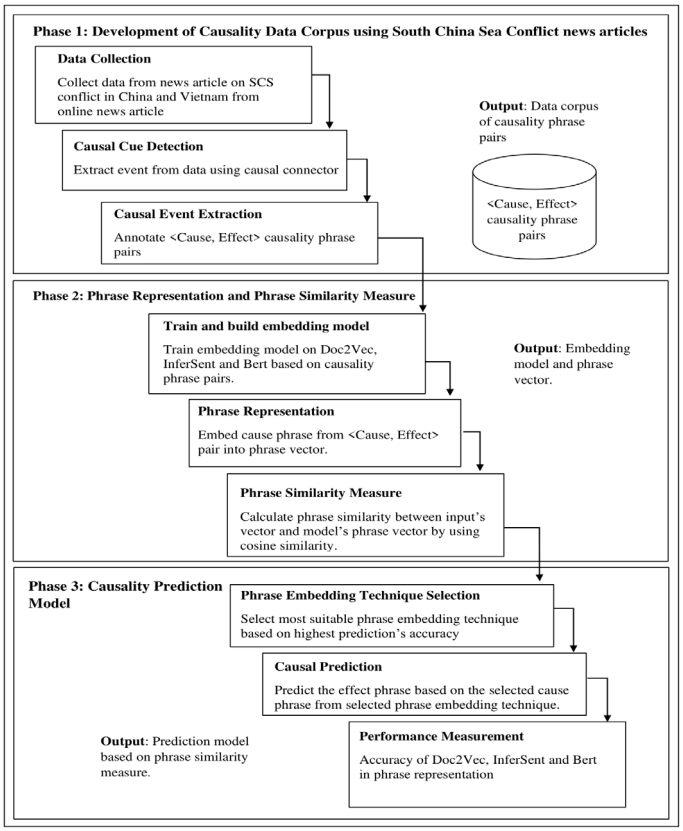

Fig. 5 Research Framework

The study is divided into three phases as shown in Figure 7. Output of Phase 1 is collection of <cause, effect> causality pairs, output of Phase 2 is embedding vector using three different embedding techniques. The best among the three embedding techniques will be selected to be implemented in Phase 3, which is the development of the proposed causal prediction model.

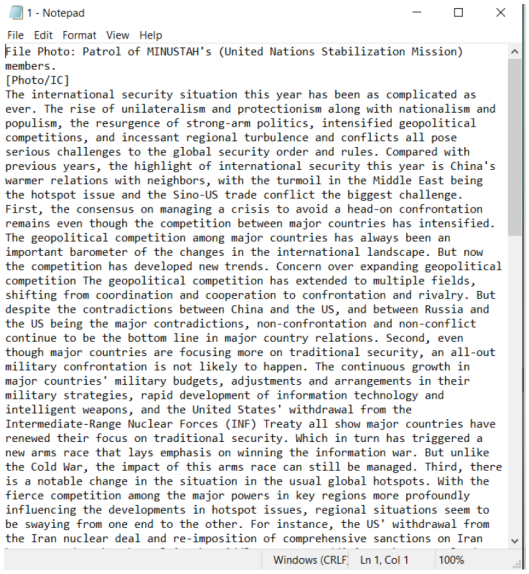

Fig. 6 Sample of news article crawled from online news

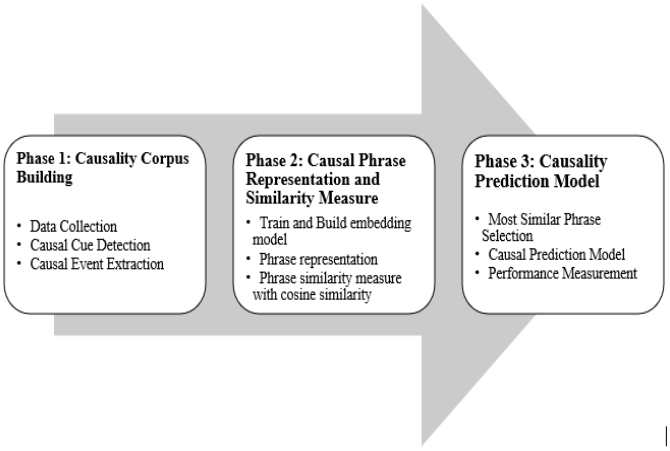

Fig. 7 Overall flow of experiment

In Phase 1, some causal news are crawled by causal cue detection. Then, Brat annotation tool is used to annotate "Cause" and "Effect" phrases and stored in an *ann* file. Figure 8 shows the conversion of *ann* file to tuple format for further processing.

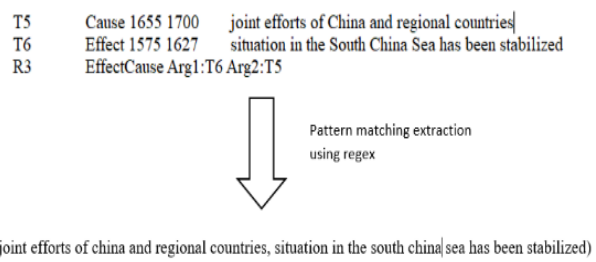

Fig. 8 Causality pair convert from ann file to tuple

In Phase 2, embedding techniques, Doc2Vec, InferSent, BERT are loaded with pre-trained models. Doc2Vec used English Wikipedia DBOW, InferSent used crawl-300d-2M-vec while BERT used bert-base-nli-mean-token. Then, each embedding technique learned the vocabulary by using training data. Testing data is treated as input queries in determining the

accuracy of the model by getting the most similar phrase, measured by cosine similarity.

In Phase 3, evaluation is done by predicting truth label and accuracy of correctly predicted events. After obtaining the result, the embedding technique with highest accuracy is selected and treated as the best embedding technique to be part of the causal prediction model.

IV. Result, Analysis and Discussion

Table 1 shows the result of predicted label on each of the test sets. Based on Table 1, BERT has the highest number of correctly predicted labels of "1" (true) which are 20 out of 31 in Test Set 1, 14 out of 31 in Test Set 2, and 60 out of 140. Table 2 shows the accuracy of each embedding technique. Table 2, tabulates the results in percentage of correctly predicted labels. Based on the results, BERT is the best-fit embedding technique. Therefore, BERT has been included in building causal prediction models. The good performance of BERT, could be due to its capability to include the contextual relation among words unlike the other two techniques. Besides, its superiority in the performance could also attributed to the usage of MASK for the bi-directional encoding.

TABLE 1 RESULT OF THE PREDICTED LABEL ON EACH TEST SET

Techniques	Test Set 1		Test Set 2		SemEval 2010 Task 8	
	0	**1**	**0**	**1**	**0**	**1**
Doc2Vec	24	7	19	12	105	35
InferSent	21	10	23	8	107	33
BERT	11	20	17	14	80	60

TABLE 2 ACCURACY OF THE EMBEDDING TECHNIQUES

Techniques	Accuracy % (Test Set 1)	Accuracy % (Test Set 2)	Accuracy % (SemEval 2010 Task 8)	Average Accuracy %
Doc2Vec	22.58	38.71	25.00	28.76
InferSent	32.26	25.81	23.57	27.21
BERT	**64.52**	**45.16**	**42.86**	**50.85**

Figure 9 shows the example of output from the proposed causal prediction model built using BERT embedding technique. For the input "China won trade war", logically the output event is related to security, political and economic crises. The predicted event shows the effect of "China won trade war" which successfully provides correct effect based on event causality.

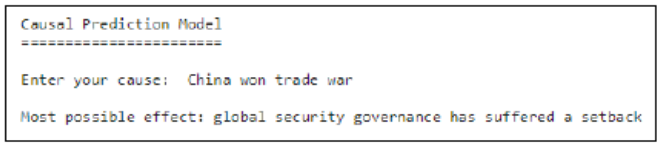

Fig. 9 Example of Causal Prediction Model

V. Conclusion

In conclusion, this study has developed a causal prediction model using the case study of SCS dispute. In the first phase, <cause, effect> causality pairs were successfully extracted by using causal cue words and annotation tools. Then, the comparison was done among these three embedding techniques (BERT, InferSent and Doc2Vec). BERT appears to be the best-fit embedding technique, and was deployed in our causal-prediction model.

Acknowledgement

We would like to thank Universiti Teknologi Malaysia and UTM RMC for funding this project under UTM TDR Grant PY/2018/03545.

References

[1] A. Fensom, "$5 Trillion Meltdown: What If China Shuts Down the South China Sea?," 16 July 2016. [Online]. Available: https://nationalinterest.org/blog/5-trillion-meltdown-what-if-china-shuts-down-the-south-china-16996.

[2] A. Ni, "Escalating Power Rivalries in the South China Sea Raise Concern," 2018. [Online]. Available: https://theglobalobservatory.org/2018/11/escalating-power-rivalries-south-china-sea-raise-concern/.

[3] G. Robin, "The Spratly Islands Dispute: International Law, Conflicting Claims, and Alternative Frameworks For Dispute Resolution," 2014.

[4] L. Zhen, "What's China's 'Nine-Dash Line' and why has it created so much tension in the South China Sea?," South China Morning Post, 2014.

[5] C. Taylor, "Structured vs. Unstructured Data," March 2018. [Online]. Available: https://www.datamation.com/big-data/structured-vs-unstructured-data.html.

[6] Y. a. S. a. Warren, "Newspaper archives+ text mining= rich sources of historical geo-spatial data," in IOP Conference Series: Earth and Environmental Science, 2016.

[7] T. a. S. Mikolov, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, 2013.

[8] M.-A. AsencioCortes, "Medium--large earthquake magnitude prediction in Tokyo with artificial neural networks," Neural Computing and Applications, vol. 28, pp. 1043--1055, 2017.

[9] K. a. M.-A. Asim, "Earthquake magnitude prediction in Hindukush region using machine learning techniques," Natural Hazards, vol. 85, pp. 471--486, 2017.

[10] T. K.-C. a. W. E. Tzu, "Mining event sequences from social media for election prediction," in Industrial Conference on Data Mining, 2016.

[11] Y. a. M. Ning, "Modeling precursors for event forecasting via nested multi-instance learning," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[12] ChinaPower, "How much trade transits the South China Sea?," 2016. [Online]. Available: https://chinapower.csis.org/much-trade-transits-south-china-sea/.

[13] L. Zhen, "What's China's 'Nine-Dash Line' and why has it created so much tension in the South China Sea?," South China Morning Post, 2014.

[14] P. Mirza, "Extracting Temporal and Causal Relations between Events," in Proceedings of the ACL 2014 Student Research Workshop, Baltimore, Maryland USA, 2014.

[15] Q. a. M. T. Le, "Distributed representations of sentences and documents," in International conference on machine learning, 2014.

[16] A. K. Conneau, "Supervised learning of universal sentence representations from natural language inference data," arXiv preprint arXiv:1705.02364, 2017.

[17] J. a. C. M.-W. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[18] F. Torabi Asr, "Querying Word Embeddings for Similarity and Relatedness," in Proceedings of the 2018 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018.

[19] T. a. S. Mikolov, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, 2013.

[20] Granroth-Wilding, M. and Clark, S. (2016). What happens next? event prediction using a compositional neural network model. In Thirtieth AI Conference on Artificial Intelligence.

[21] Preethi, P. G., Uma, V. et al. (2015). Temporal sentiment analysis and causal rules extraction from tweets for event prediction. Procedia computer science. 48, 84–89.

[22] Brat rapid annotation tool. (n.d.). Retrieved January 16, 2021, from https://brat.nlplab.org/