DOI: 10.37943/TSYV3590

A. Mussina
PhD student of Computer Science, Department of Computer Science
mussina.aigerim95@gmail.com, orcid.org/0000-0002-7043-0810
Al-Farabi Kazakh National University, Kazakhstan

S. Aubakirov
PhD, Department of Computer Science
aubakirov.sanzhar@gmail.com, orcid.org/0000-0002-8416-527X
Al-Farabi Kazakh National University, Kazakhstan

P. Trigo
PhD, Professor Adjunto, Department of Electronics and Telecommunications and Computer Engineering
paulo.trigo@isel.pt, orcid.org/0000-0001-5850-615X
ISEL - Instituto Superior de Engenharia de Lisboa; GulAA; LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

PARAMETRIZED EVENT ANALYSIS FROM SOCIAL NETWORKS

Abstract: The growth of data in social networks facilitate demand for data analysis. The field of event detection is of increasing interest to researchers. Events from real life are actively discussed in the virtual space. Event detection results can be used in a variety of applications, from digital marketing to collecting data about natural disasters. Thereby, researchers face the emergence of new algorithms along with the improvement of existing solutions in the event detection field. This paper proposes improvements to the SEDTWik (Segment-based Event Detection from Tweets using Wikipedia) algorithm. The SEDTWik algorithm is designed to detect events without contextual guidance. The overall SEDTWik detection process excludes the perspective of a topic, or multi-topic, guided (or semi-supervised) event detection approach. As a result, some interesting narrowly focused events are not detected as they are weakly relevant in a broader context (e.g., Wikipedia) although acquiring relevance within a conditioned context. Therefore, there is a need for an adaptive perspective where data is to be analysed against a set of narrower topics of interest. This paper shows that SEDTWik gains expressive power after being extended with multi-topic semi-supervision. The evaluation of the current proposal uses the well-known corpora with labeled events, Events2012. In the Events2012 dataset used notation category for events, meaning that events are combined by a certain topic. SEDTWik with topic dictionaries was checked across all categories. In the main part of the article, it is also explained the process of topic dictionary construction from Events2012 labeled tweets. At this stage of the research, in all tasks unigrams were used. SEDTWik with dictionaries showed improved accuracy, and more events were found within a certain category.

Keywords: event-detection with multi-topic semi-supervision, SEDTWik, social media, dictionary, Events2012.

Introduction

Now, social networks are strongly bound to the daily lives of people. First of all, news is derived from social networks, and people share the world around us with other people through social media. It has become commonplace for people to post every moment of their lives

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 10, JUNE 2022

on the Internet. At the moment, there are many different approaches for analysing big data, including unstructured data from social networks. One of the analysis processes is called event detection, and its purpose is to collect the most discussed topics and extract their features.

The process of social network analysis involves working with continuously growing data. In the previous work, we have already presented our proposal for a data crawling and pre-processing system [1]. The current work presents proceeding of data analysis for event detection by improving the SEDTWik (Segment-based Event Detection from Tweets using Wikipedia) algorithm [2]. The SEDTWik is a state-of-the-art algorithm based on the Twevent previous technique [3].

The SEDTWik algorithm consists of four steps. In the first step, it gets segments from tweets. In the second step, it calculates the "bursty" value of a segment and continues the calculation with bursty segments. In the third step, it constructs clusters of bursty segments. Finally, SEDTWik gives an event summary with the top segments. The first and third steps, tweet segmentation and bursty segment clustering, algorithm uses Wikipedia as one of the input data. In the main part of the article, the use of Wikipedia in the algorithm will be described in more detail. At the moment, it is worth noting that Wikipedia is a fairly extensive dataset of words on a variety of topics. In this regard, the detection of events within the same time period includes the most discussed events outside a certain focus. Here is described a different perspective as it is assumed that event detection should contemplate the notion of relevance that is assigned by users according to their topic of interest. For example, it would be convenient for various TV awards to follow the opinion of the public on social networks in the context of their activities without other events' noise.

Given this motivation, we propose an extension of the algorithm to replace the use of Wikipedia by the inclusion of multiple topics (multi-topic) where each topic is represented by a dictionary-like structure. Although it was not yet fully obliterated the Wikipedia usage, it has been replaced in the major places. Those major places are tweet segmentation and segment bursty extraction. Initially, segments were extracted according to Wikipedia titles. The bursty of a segment is calculated via text probability derived from Wikipedia.

Twitter is one of the biggest online social media. It is usual to find research about Event Detection based on data from Twitter [2-5]. Furthermore, it has advanced developer APIs available through a developer account. Tweets were crawled using APIs and analysed. The evaluation of the current proposal uses the well-known corpora with labeled events, Events2012, which consists of tweet ids from 10/10/2012 to 07/11/2012 [4]. Due to the confidentiality of the data, the dataset does not contain the tweets themselves in text form, but their id. At the moment, some tweets have been deleted, or the user has closed public access to his/her page. Therefore, detected events may differ from those received by the authors of the article. This article presented experiments include comparison of the events detected through SEDTWik in its original form with the events detected using SEDTWik with topic dictionaries. The results obtained indicate that the use of topic dictionaries improve event detection in the category under consideration.

Related works

Event detection process has been applied for various online social networks (OSN) and purposes. Some construction of an event model depends on OSN structure. For example, in Twitter researchers take into account retweets, user's count of followers [2, 6].

At the beginning, one of the common interest of researchers is the natural disasters. Since natural disasters are not frequent events, one work concentrated on the evolution of rare event, like storm, in the real world by analyzing activities in virtual world [7]. Authors denotes the idea for social media activities temporal pattern constriction. Another work, a survey, on

event detection techniques for natural disaster events, trending topics and public opinion events identified more specific social networks, like newswire, web forums, emails, blogs and microblogs [8]. Authors defined that domain dependence is a huge challenge for researchers because techniques are extremely situational dependent.

Nowadays there are researches contributing in healthcare sectors. Online information from Weibo is used to detect foodborne diseases [9]. Authors preprocessed data as a vector of extracted features. Finally Support Vector Machine was used to construct a model for event detection.

During this research it was favoured the articles that have publicly available programming code [10]. This simplifies the verification of research results and helps to understand whether it would be possible to apply each algorithm to crawled data.

One of the idea is to find "explicit event descriptors" by answering questions. For example, from the journalist's practice it could be taken 5Ws with 1H: What? Who? Where? When? Why? and How? [11]. Questions also could be almost natural [12]. However, since OSN messages usually are short it is hard to answer all questions. Most likely, such algorithms will be used for full web articles.

In another work, the authors pay attention to the extraction of several events from one sentence. [13]. The fact that the work uses texts as small as sentences suggests that it can be used on data from Online Social Networks (OSN). Nevertheless, the code was difficult to reproduce as it is hard to get it up and executing. Also the dataset ACE 2005 is not free which makes it too hard to explore.

The vision of an event emerging in a social media (network) in this work is that it is likely to be a burst of messages related to a certain topic, time and place. Close to the vision of this work, the article detects an event from identifying segments taken from tweets and proposes the SEDTWik algorithm, which the authors have made public. [2]. The SEDTWik core idea is that a word or phrase from a tweet text becomes a segment if it is in the Wikipedia Titles Dataset [14]. The most discussed segments are called bursty segments, but segment frequency is not the only metric and the process also takes into account the user who posted the text, the hashtags and the presence of the user mention. This work was accurately written with explanations. The SEDTWik algorithm reveals acceptable results when compared with previous Twevent work.

Main part

In this section the changes in SEDTWik are described and mentioned what was left intact. At first it is presented the SEDTWik overall workflow and then formulated proposed SEDTWik extension.

The SEDTWik workflow goes through the following processes: a) tweet segmentation, b) bursty segment extraction, and c) bursty segment clustering. As it was mentioned above in the Introduction, the Wikipedia data has a role in two of those processes, tweet segmentation and bursty segment clustering. The Wikipedia Titles Dataset is used in tweet segmentation because if a word or phrase from a tweet exists in the title dataset then it is considered to be a segment. The bursty segment calculation uses the expected probability that the segment will appear in the tweet. The expected probability is calculated on the base of Events2012 corpora, which is also used for SEDTWik evaluation. The clustering process uses the fraction of frequency of segments in line with tf-idf similarity of the set of tweets. It is called as similarity function. During this clustering process the Wikipedia Keyphraseness dataset is used in the newsworthiness calculation. The Wikipedia Keyphraseness dataset is a set of probability that a word or phrase will appear in the article as an anchor text [15]. The anchor is an HTML tag, means the text has hyperlinks to another article.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 10, JUNE 2022

The first SEDTWik test on the currently available tweets from the Events2012 corpora resulted in 1 to 5 number of events of different categories during each day from 10/10/2012 to 07/11/2012. The segmentation of the tweet is based on the Wikipedia Titles Dataset, therefore segment variations are limited to titles only. Then SEDTWik calculates the segment bursty using the probability taken from the Events2012 corpora. From the above statements, it can be concluded that the variety of detected events depends on the initially identified segments and their expected probability which are highly related to the broad data provided by the Wikipedia.

In this research work Wikipedia titles and segment probability were replaced, respectively, with topic dictionaries and thematicity value. The thematicity value is a coefficient that denotes the degree of belonging of a word or phrase to the topic dictionary; the higher the value, the more the word relates to the dictionary.

The next sections describe topic dictionaries extraction and substitution of Wikipedia with topic dictionaries in SEDTWik.

Topic dictionary

In the Events2012 dataset used notation category for events, meaning that events are combined by certain topic. In this paper 'category' notation used towards corpora text and 'topic' notation towards dictionary. So certain topic-dictionary refers to certain event category tweets.

In our previous work, we had already performed the extraction of a topic dictionary of words related to the scenario of emergency situations [16]. The idea is that words specific to a certain category of events may appear in the texts of other categories, but most often they will occur in their own category. The methodology for such dictionary extraction [17] needs two corpora: a) the target corpora that contains text about the topic of interest and b) the common corpora that contains mixed topics. In this research stage only unigrams are used in dictionaries because current version of the architecture for event detection only calculates unigram frequency.

As an example, let's consider that it is needed to build a topic dictionary for "Sports". The target corpora should consist of text about sport only. The common corpora should contain text of other topics. If the word 'cycling' appears in both corpora and if the frequency in the target corpora is higher than in the common corpora, then it could be concluded that 'cycling' is a word about "Sports".

The process of thematicity calculation and including word to the topic dictionary is based on the expression 1. The M_w is the thematicity coefficient and it is calculated only if the word w occurs in both target and common corpora. The N_w^{target} is the frequency of word w in the target corpora. The N_w^{common} is the frequency of word w in the common corpora. From the (1) it is known that if the word w occurs more often in the target corpora, then M_w is a positive value. On the other case M_w will be negative value representing that word w may occur in the target corpora, but it is considered more as a common word.

$$M_w = log \frac{N_w^{target}}{N_w^{common}} \quad (1)$$

SEDTWik with topic dictionaries

The simplified form of the SEDTWik algorithm is presented in a Fig. 1. In this work it is considered that SEDTWik is a "black box" with its own formulation for bursty extraction and clustering. The overall process accepts two inputs:

a) a data stream from OSNs (Online Social Networks) such as Twitter and Telegram

b) the Wikipedia Title Dataset and the Wikipedia segment probability.

The final version of the SEDTWik with topic dictionaries is presented in the Fig. 2. Wikipedia input was replaced with the topic dictionaries. At the step of extracting segments, the segments

DOI: 10.37943/TSYV3590

corresponding to the topic dictionary will be extracted. At the segment bursty calculation the thematicity value of segment from the topic dictionary will be used. As a result, segments will be focused on the selected topic and SEDTWik will find the most relevant events according to the available topics of interest.

The Wikipedia only stayed in keyphraseness values, which is used during event newsworthiness calculation. At this stage of the study, it is not valueable, since it does not affect the event detection itself.

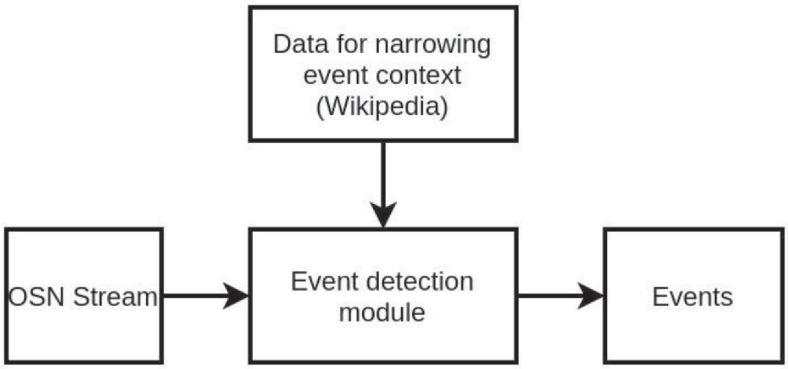

Figure 1. Simplified SEDTWik

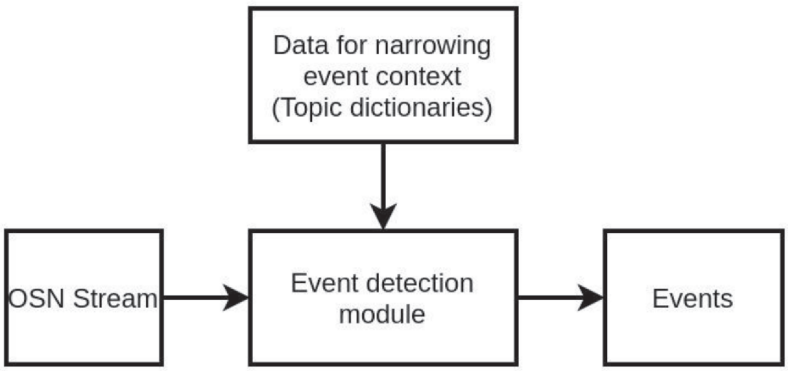

Figure 2. Simplified SEDTWik with topic dictionary

Results

The tests were conducted using Events2012 corpora which consists of tweet ids, event code and 8 event categories. Each tweet is labeled as one of 506 events and each event has a corresponding category. For example, events connected to the hostilities and terrorist attacks combined to the "Armed Conflicts & Attacks" category, events about celebrities and movies collected in the "Arts, Culture & Entertainment", and so on. The time period of tweets is from 10/10/2012 to 07/11/2012. Since this research work needed a tweet text and other metadata the public Twitter API was used with guest tokens and tweet ids. The test data has 152 952 labeled tweet ids, but only 63 271 tweets are currently available for public API usage. Some tweets were deleted, and some tweets belong to private accounts. In this section the result of topic dictionary extraction and SEDTWik improvements via topic dictionaries are presented.

Topic dictionary

Corpora Events2012 has 8 categories of events: "Armed Conflicts & Attacks", "Arts, Culture & Entertainment", "Business & Economy", "Disasters & Accidents", "Law, Politics & Scandals", "Miscellaneous", "Science & Technology", "Sports".

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 10, JUNE 2022

Topic dictionary was extracted for each event according to (1). The target corpora is a set of all tweet text from target event. The common corpora is a set of all tweet text from other events. For example, in the Table 1 the top-5 words from three categories with their thematicity value are presented.

Indeed, words in a dictionary describe the name of the topic dictionary. The higher the thematicity value, the more the word applies to the topic.

Table 1. Top-5 words from topic dictionaries

Armed conflicts & Attacks		Disaster & Accidents		Arts, Culture & Entertainment	
word	**thematicity**	**word**	**thematicity**	**word**	**thematicity**
bombers	4.443	superstorm	4.543	bollywood	4.564
philippine	4.263	shark	4.538	daniel	4.518
qaeda	4.135	riyadh	4.234	premiere	4.430
rockets	4.094	arabia	3.970	skyfall	4.430
rebels	4.074	floods	3.412	omarion	4.369

In the Table 1, the words describe the events that took place in the existing period. For example, in the October 26, 2012, the premiere of the film «Skyfall» took place. Words about this premiere are in the topic dictionary «Arts, Culture & Entertainment».

SEDTWik with topic dictionaries

The SEDTWik algorithm was used as provided by the authors from github [18]. The Wikipedia Titles Dataset and expected segment probability were substituted with constructed topic dictionaries. The maximum segment length was 4 word. In this research work entities-only mode was not used. Also the use-retweet-count and use-followers-count were turned false. Number of neighbours was 3.

SEDTWik with topic dictionaries was checked across all categories. Since corpora is labeled, the accuracy was calculated for each category. The accuracy is calculated as a ratio from dividing the number of events of certain category, as determined by the algorithm, by the total number of events in this category for the entire period of time. The accuracy is presented in Fig. 3, where along X-axis: 1 – Armed Conflicts & Attacks, 2 – Arts Culture & Entertainment, 3 – Business & Economy, 4 – Disasters & Accidents, 5 – Law Politics & Scandals, 6 – Miscellaneous, 7 – Science & Technology, 8 – Sports.

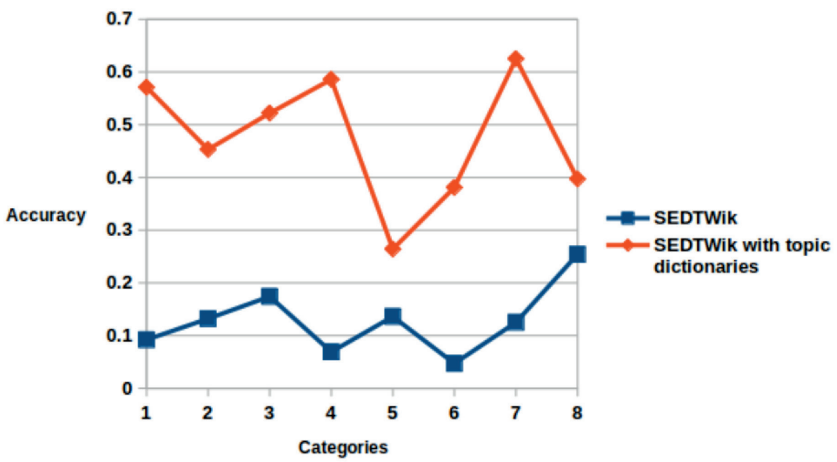

Figure 3. SEDTWik and SEDTWik with topic dictionaries accuracy

DOI: 10.37943/TSYV3590

The value of thematicity concept is not constrained to the probability range (from 0 to 1) so it was normalized, such that the thematicity values fitted the SEDTWik formulation. Thematicity was normalized by the maximum observed value multiplied by 1000. Here the scaling was applied to satisfy the sigmoid function values within SEDTWik formulation. The expected segment probability from Wikipedia has a degree usually of 10^{-8}. However, using a degree of 10^{-3} also resolved issue with sigmoid function errors.

In consonance with results it is clear that using dictionary representation of topics of interest it is possinle to improve event detection from bursty messages.

Discussions

Through the use of topic dictionaries, an improvement in the number of detected events in a particular event category was achieved. It turns out that if a user of proposed event detection system wants to know what events of an armed or conflict nature occurred over a certain period of time, then he/she will need to indicate the appropriate dictionary. The users will be more likely to find the events they need. For example, on October 10, 2012 events detected by SEDTWik are about 1) Music Awards and 2) about two American scientists, Robert Lefkowitz and Brian Kobilka, who win the 2012 Nobel Prize. Using "Armed conflicts & Attacks" topic dictionary, user can find events about 1) a suicide bomb in Damascus, 2) Heriberto Lazcano Lazcano, the top leader of the criminal organization Los Zetas, was killed, 3) Malala Yousafzai, a 14 year old activist for women's education rights is shot by Taliban gunmen in the Swat Valley. However, during the test with topic dictionaries the words were unigrams and in case of SEDTWik parameters given by default were used, meaning maximum length of segment was 4. In the future we plan to increase number of N-grams in dictionaries.

Conclusion

In this paper the state-of-the-art algorithm, SEDTWik, for event detection was analysed and improved. Starting from the idea that people need to detect events according to some topics of interest to them (ToI), it was proposed to replace the extensive and general Wikipedia data with ToI-oriented dictionaries. The topic dictionaries using labeled tweets by categories of events were constructed. The accuracy of event detection in a context of focused search was improved. Valuable events in each category are not missed. However, during the clustering process (used in SEDTWik) the newsworthiness calculated according to the Wikipedia keyphraseness dataset.

This research study can be further developed within the framework of adaptive event analysis, where the topic of events is a variable parameter and does not limit the scope of event detection. The importance of research lies in improving the accuracy of event identification through focused search.

In the future work we intend to continue research in three approaches: 1) represent event as a vector, 2) adapt methods for real-time event detection and 3) detect associations between events. In the first upcoming task, outcome events would be represented as vectors with additional information, like sentiment, location, time. According to the second future task, the software should be adaptive and dynamic. It should process a new ToI in real-time over streaming data. The third future task will show dependencies between events and in a perspective will help in event prediction task based on historical data.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 10, JUNE 2022

References

1. Mussina, A.B., Aubakirov, S.S., & Trigo, P. (2021). An Architecture for Real-Time Massive Data Extraction from Social Media. Communications in Computer and Information Science, 138–145. https://doi.org/10.1007/978-3-030-78759-2_11
2. Morabia, K., Bhanu Murthy, N. L., Malapati, A., & Samant, S. (2019). SEDTWik: segmentation-based event detection from tweets using Wikipedia. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, 77–85. https://doi.org/10.18653/v1/n19-3011
3. Li, C., Sun, A., & Datta, A. (2012). Twevent: segment-based event detection from tweets. Proceedings of the 21st ACM International Conference on Information and Knowledge Management - CIKM '12, 155–164. https://doi.org/10.1145/2396761.2396785
4. McMinn, A.J., Moshfeghi, Y., & Jose, J.M. (2013). Building a large-scale corpus for evaluating event detection on twitter. Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management – CIKM '13, 409–418. https://doi.org/10.1145/2505515.2505695
5. Bekoulis, G., Deleu, J., Demeester, T. & Develder, C. (2019). Sub-event detection from twitter streams as a sequence labeling problem. arXiv preprint arXiv:1903.05396
6. Chen, X., Zhou, X., Sellis, T., & Li, X. (2018). Social event detection with retweeting behavior correlation. *Expert Systems with Applications*, *114*, 516–523. https://doi.org/10.1016/j.eswa.2018.08.022
7. Lu, X. S., Zhou, M., Qi, L., & Liu, H. (2019). Clustering-Algorithm-Based Rare-Event Evolution Analysis via Social Media Data. *IEEE Transactions on Computational Social Systems*, *6*(2), 301–310. https://doi.org/10.1109/tcss.2019.2898774
8. Goswami, A., & Kumar, A. (2016). A survey of event detection techniques in online social networks. *Social Network Analysis and Mining*, *6*(1). https://doi.org/10.1007/s13278-016-0414-1
9. Cui, W., Wang, P., Du, Y., Chen, X., Guo, D., Li, J., & Zhou, Y. (2017). An algorithm for event detection based on social media data. *Neurocomputing*, *254*, 53–58. https://doi.org/10.1016/j.neucom.2016.09.127
10. Papers with Code - The latest in Machine Learning. (2021, August 25). Papers with Code. Retrieved August 25, 2021, from https://paperswithcode.com/
11. Hamborg, F., Breitinger, C. & Gipp, B. (2019). Giveme5w1h: A universal system for extracting main events from news articles. arXiv preprint arXiv:1909.02766
12. Du, X. & Cardie, C. (2020). Event extraction by answering (almost) natural questions. arXiv preprint arXiv:2004.13625
13. Liu, X., Luo, Z. & Huang, H. (2018). Jointly multiple events extraction via attention-based graph information aggregation. arXiv preprint arXiv:1809.09078.
14. ENwiki-latest-all-titles. (2021). Wikimedia Downloads. Retrieved August 26, 2021, from http://dumps.wikimedia.org/enwiki/latest/enwiki-latest-all-titles-in-ns0.gz
15. Wikipedia Keyphraseness. (2021). Aixin's Homepage. Retrieved August 26, 2021, from https://personal.ntu.edu.sg/axsun/datasets.html
16. Mussina, A. & Aubakirov, S. (2017) Dictionary extraction based on statistical data. KazNU Bulletin. Mathematics, Mechanics, Computer Science Series, 94(2), 72–82.
17. Barr, I. (2016, April 20). Heavy Metal and Natural Language Processing - Part 1. Degenerate State. Retrieved September 20, 2016, from http://www.degeneratestate.org/posts/2016/Apr/20/heavy-metal-and-natural-language-processing-part-1/
18. SEDTWik-Event-Detection-from-Tweets. (2020, July 13). Github. Retrieved August 26, 2021, from https://github.com/kevalmorabia97/SEDTWik-Event-Detection-from-Tweets