The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 2021

Volume 16, Pages 82-86

IConTES 2021: International Conference on Technology, Engineering and Science

Evaluation of an Event Detection Algorithm for Russian and Kazakh Languages

Aigerim MUSSINA
Al-Farabi Kazakh National University

Sanzhar AUBAKIROV
Al-Farabi Kazakh National University

Paulo TRIGO
Universidade de Lisboa

Abstract: The Event Detection area is gaining increasing interest among researchers. The social media data growth induces the emergence of new algorithms along with the improvement of existing solutions. In this paper we propose to improve of existing algorithm for event detection, SEDTWik (Segment-based Event Detection from Tweets using Wikipedia). The authors define event as a set of similar segments of words within a given time window. A segment is defined as a word or phrase taken from the analyzed text data. The SEDTWik uses Wikipedia as a "supervisor" to identify the segments, to calculate the segments' bursty value and to calculate the segments' newsworthiness. We examined the SEDTWik algorithm using our data from Telegram online social network. The overall network message construction of Twitter is different from that of Telegram. Therefore, we transformed the Telegram meta-data to fit the SEDTWik requirements. Another much relevant difference in our experiment lies in the fact that our corpora contain messages in Russian and Kazakh languages. Our results show that the SEDTWik algorithm is strongly dependent on the broad and unfocused Wikipedia data. Such dependency was shown to have a loss effect on the event detection accuracy. This result founds our motivation to improve the SEDTWik algorithm using dynamically calculated segment probabilities from the analyzing data streams.

Keywords: Event detection, SEDTWik, Russian language, Kazakh language

Introduction

Social data analysis is gaining in importance due to the increasing use of social media by people. One of the social data analysis area is Event Detection. In this paper we present our initial and current steps in the field of Event Detection analysis. Our initial step was driven by the need to gather data from social media, so our previous work (Mussina, 2021) builds an architecture for data crawling from Online Social Networks (OSN). In our current step we are driven by the need to add an effective and accurate Event Detection module to our (already implemented) architecture.

We have chosen Telegram OSN as our first data source (Mussina, 2021) given that in January 2021 Telegram's world-wide monthly active users reached 500 million users. The Telegram is a popular social network among the countries from Commonwealth of Independent States (CIS). During our previous experiments we analyzed locally popular channels and groups that use the Russian and Kazakh languages. From those previous experiments we proceed to current work where we are focused on applying the Event Detection algorithm to the Russian and Kazakh language data.

During this research stage we examined several existing Event Detection algorithms. In the end we chose the work (Morabia, 2019) which presented the algorithm called SEDTWik (Segment-based Event Detection from Tweets using Wikipedia). In this algorithm the data stream divided by user-defined time window and goes as input data. SEDTWik starts with tweet text segmentation. After that the bursty value extracts from segment according to its expected probability and other meta-data. Segments with the highest value clustered to events. As a result, each event is characterized by time window, segments and newsworthiness. All details we describe in next section.

Related Works

During this research we explored the articles that provide publicly available code (Papers with code). Therefore, it was easy to us to reproduce the algorithms and understand whether it would be possible to apply each algorithm to our data. Several algorithms were tested on their test data.

The works (Hamborg, 2019; Du, 2020) are based on the idea of "answering questions". In Hamborg (2019) authors used 5W1H questions: What? Who? When? Where? Why and how? Another researcher (Du, 2020) used "almost natural questions". However, here we have many questions and small OSN messages could not give an answer to all of them. The work (Liu, 2018) suggests to extract multiple events from a single sentence. The data partition as a single sentence is similar to our OSN messages. However, the code was difficult to run and the explored dataset ACE 2005 is not free.

Finally, we examined the SEDTWik algorithm presented in a clearly written paper with, explaining formulas and illustrating examples. The code was easy to execute and test on small part of the Events2012 dataset (McMinn, 2013). The base idea of the algorithm is the usage of Wikipedia. Tweet is segmented by phrases and if segment is in the Wikipedia Titles Dataset (Wikimedia Downloads) then segment goes further. In the next step the bursty value is extracted using segment expected probability, user followers count, retweet count. The expected probability denotes the probability that a segment will appear in the text of a tweet. It is calculated on the dataset Events2012. Next segments with the highest bursty values clustered and generated event summary with newsworthiness. The newsworthiness calculated from Wikipedia Keyphraseness (Sun). Therefore, SEDTWik test on our data should include the substitution of Wikipedia English data to Russian and Kazakh data.

Method

In this section we present the adaptation of SEDTWik algorithm to fit the data from Telegram OSN. The main changes considered base programming model of OSN data and expected segment probability.

Table 1. Comparison of Twitter and Telegram

Twitter	Telegram
Users should be registered. Messages are small. Can use hashtags (#) and user mentioning (@).	
User can post tweet on his/her page	User should be a member of a group/channel to post a message in group's chat and in threads of channels
User can write a comment under another user's tweet	User can reply to all messages inside a group
User can retweet another user's tweet	User can forward another user's message
Users have own pages	Users can enter the groups, channels

Telegram Model

The Telegram and Twitter are different OSNs. The Telegram acts like messenger with extended opportunities. The Twitter is more like a place for self-expression and interaction with all users inside the network. We denoted common and different aspects of each network (Table 1). One of the main differences is that Telegram

has groups, super groups and channels. The member's capacity and administrators' opportunities are growing from groups to channels.

The SEDTWik data model includes tweet's text, creation date, user with its id and followers count, retweet count and entities. The entity is a hashtag and/or user mention inside tweet. Therefore, it doesn't have some functionalities presented in Twitter. For example, Telegram doesn't have retweets, user followers, and hashtags are available, but they are not usually used. We substituted some Twitter specific fields in a model to bring Telegram model (Figure 1). As depicted in the Figure 1, we substituted: a) followers count by count of chat/channel members, and b) retweet count by count of messages in thread.

Twitter	Telegram
{ "created_at": "2021-10-25 15:00:00", "text": "@Random, simple text #test", "user": { "id": 000000000, "followers_count": 118 }, "retweet_count": 100, "entities": { "hashtags": ["test"], "user_mentions": ["Random"] } }	{ "created_at": "2021-10-25 15:00:00", "text": "@Random, simple text #test", "user": { "id": 000000000, "chat_members_count": 118 }, "thread_message_count": 100, "entities": { "hashtags": ["test"], "user_mentions": ["Random"] } }

Figure 1. Twitter and Telegram JSON models

The authors of SEDTWik stated that tweets from users with millions of followers should have more weight than tweets from users with few followers. We transposed that idea to Telegram and propose that messages from chats/channels with thousands of members should have higher weight than those from chats/channels with few members.

The retweet transposition (from Twitter to Telegram) is a difficult task in terms of compliance. At first, we thought that message forwarding is similar to retweet process. Both the concepts of retweet and of forwarding a message are responsible for spreading information among members of groups or channels and retweets do the same. However, in Telegram it is difficult to know the amount of message forwarding. We only know that message was forwarded, but we don't know how many times. Another solution is to use Telegram threads that are a separate branch of conversation in channels and super chat groups. The number of messages in a thread may be used as a metric on the users' interest in a thread's discussion. Such a metric is similar to the retweet count metric, because retweet count also shows how many people are interested in a tweet.

English Wikipedia Substitution with Russian and Kazakh

Firstly, we need to substitute the Wikipedia Titles Dataset with publicly available Wikipedia dumps. We downloaded (Wikimedia Downloads) Wikipedia Titles Dataset for Russian and Kazakh languages. Secondly, we substituted the expected probabilities, we used our tokens frequency and calculated probability. Our tokens frequency calculated during text preprocessing. In our architecture, after a message is crawled it goes to the text preprocessing tokenization and token frequency calculation processes.

Thirdly, we need to substitute the Wikipedia Keyphraseness file, however it is hard to replace file quickly, because it was calculated separately and the Wikipedia articles dump creation date was January 30, 2010. So the data even for English data could be old. Since the Keyphraseness demonstrate the probability that segment could be anchor of Wikipedia article, we could replace it via our dictionary. The dictionary is a set of words with special thematicity value. Word is present in a dictionary if it is connected to the dictionary topic. For example, we took sport-topic dictionary. All words inside this dictionary describe the sport topic. The thematicity value denotes the degree of belonging to the dictionary. More details about dictionary extraction is given in our previous work (Mussina, 2017).

Results and Discussions

For the first tests, we subjectively chose the events that took place during the Olympic Games Tokyo 2020, from July 23 to August 8, 2021. The included events and the Event Detection results with Russian Wikipedia and our sport dictionary are presented below (Table 2). The event describing segments are translated to English from Russian and Kazakh for clarity in column 5. We created a sport dictionary based on two types of channels: sport and news. As a result, dictionary has actual information in period from 2021-07-01 to 2021-07-31.

Table 2. Sport event detection

	Event description	Date	Detected events	Detected events. Translation
1	Eldos Smetov won first medal for Kazakhstan in judo	2021-07-25	EVENT 1: игорь, сон, медаль, EVENT 2: нашим, состав, спорт,	EVENT 1: igor, son, medal, EVENT 2: our, composition, sport,
2	Igor Son won second medal for Kazakhstan in barbell			
3	Zulfia Chinshanlo third medal for Kazakhstan in barbell	2021-07-26	EVENT 1: бокс, алтын, олимпиада, EVENT 2: медаль, live, html, чиншанло,	EVENT 1: boxing, altyn, olympics, EVENT 2: medal, live, html, chinshanlo,
4	Vasiliy Levit losing in boxing	2021-07-27	EVENT 1: баландин, теннис, мугуруса,	EVENT 1: balandin, tennis, mugurus,
5	Strong tennis match Rybakina/Bencic in ½	2021-07-29	EVENT 1: рыбакина, бенчич, арна	EVENT 1: rybakina, benchich, arna,
6	Tennis match Rybakina/Svitolina in ½ for bronze	2021-07-31	EVENT 1: финале, победы, ошибки, EVENT 2: наши, медали, турция,	EVENT 1: finals, victories, mistakes, EVENT 2: our, medals, turkey,
7	Kamshybek Kunabayev losing in boxing	2021-08-04	EVENT 1: золото, равно, бокс, EVENT 2: завтра, мог, шанс,	EVENT 1: gold, equals, boxing, EVENT 2: tomorrow, might, chance

We can see that 3 out of the 7 events were detected (i.e., an accuracy of around 40%). Therefore, we consider these, lower than 50%, accuracy as foundational for the motivation to improve results. We propose that SEDTWik algorithm could be improved via Wikipedia substitution with topic dictionary. The implementation of topic dictionary could improve results and make them more focused. Nevertheless, topic dictionary could be dynamically constructed on the basis of interested data stream.

Conclusion

In this paper we briefly present the start of our research in the field of Event Detection. We examined the SEDTWik algorithm and exploited its applicability in our corpora. We substituted the English Wikipedia Titles Dataset with the Russian and Kazakh Wikipedia Titles Dataset, used our calculated token probability and used sport topic dictionary to cluster the events. Although the results on Event Detection are often subjective after our experiments, we can conclude that SEDTWik algorithm is applicable to our data. We also expect that changing the Wikipedia usage also could improve the results. In future work we will explore substituting the Wikipedia data with dynamically created topic dictionaries.

Scientific Ethics Declaration

The authors declare that the scientific ethical and legal responsibility of this article published in EPSTEM journal belongs to the authors.

References

Du, X., & Cardie, C. (2020). Event extraction by answering (almost) natural questions. *arXiv preprint arXiv:2004.13625*.

Hamborg, F., Breitinger, C., & Gipp, B. (2019). Giveme5w1h: A universal system for extracting main events from news articles. *arXiv preprint arXiv:1909.02766*.

Liu, X., Luo, Z., & Huang, H. (2018). Jointly multiple events extraction via attention-based graph information aggregation. *arXiv preprint arXiv:1809.09078*.

McMinn, A. J., Moshfeghi, Y., & Jose, J. M. (2013, October). Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 409-418).

Morabia, K., Murthy, N. L. B., Malapati, A., & Samant, S. (2019, June). SEDTWik: segmentation-based event detection from tweets using Wikipedia. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop* (pp. 77-85).

Mussina, A. B., Aubakirov, S. S., & Trigo, P. (2020, November). An Architecture for Real-Time Massive Data Extraction from Social Media. In *International Conference on Mathematical Modeling and Supercomputer Technologies* (pp. 138-145). Springer. https://doi.org/10.1007/978-3-030-78759-2_11

Mussina, A., & Aubakirov, S. (2018). Dictionary extraction based on statistical data. *Вестник КазНУ. Серия математика, механика, информатика*, *94*(2), 72-82.

Papers with code. (n.d.). *Event detection*. Papers with code. https://paperswithcode.com/task/event-detection

Sun Aixin. (n.d.). *Wikipedia Keyphraseness*. https://personal.ntu.edu.sg/axsun/datasets.html.

Wikimedia Downloads. (n.d.). Data downloads. https://dumps.wikimedia.org/

Author Information

Aigerim MUSSINA
al-Farabi KazakhNational University
71 al-Farabi Ave., Almaty, Republic of Kazakhstan, 050040
Contact e-mail: *mussina.aigerim95@gmail.com*

Sanzhar AUBAKIROV
al-Farabi KazakhNational University
71 al-Farabi Ave., Almaty, Republic of Kazakhstan, 050040

Paulo TRIGO
ISEL - Instituto Superior de Engenharia de Lisboa; GulAA; LASIGE, Faculdade de Ciências, Universidade de Lisboa, 1959-007, Lisbon, Portugal

To cite this article:

Mussina, A., Aubakirov, S. & Trigo, P. (2021). Evaluation of an event detection algorithm for Russian and Kazakh languages. *The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 16*, 82-86.