Event extraction from Tweets

Natural Language Processing Project

URGYAN NURBOO (IIT2020172)

Github Link:

https://github.com/urgyannurboo/IIT2020172_NLP_PROJECT.git

I. ABSTRACT

The increasing use of Twitter as a platform for real-time information sharing has led to the need for effective event extraction from tweets. Extracting events from the noisy and unstructured nature of tweets presents challenges that need to be addressed. In this report, we propose a methodology for event extraction from tweets, which involves data collection, data pre-processing, event detection, and event classification. We outline the steps involved in detail and highlight the potential applications and impact of our proposed methodology in social media analytics.

II. INTRODUCTION

Twitter is a popular social media platform where users share their opinions, thoughts, and news in real-time. With over 330 million active monthly users, Twitter has become a valuable source of information for various applications such as news monitoring, disaster management, and marketing analysis. However, extracting relevant information from Twitter is challenging due to the high volume, velocity, and variety of data.

Event extraction from Twitter involves identifying and extracting events from tweets, along with their relevant attributes such as type, location, and time. This task is important for various applications such as tracking disasters, monitoring news, and analyzing social trends. However, event extraction from Twitter is challenging due to the informal nature of the language used in tweets, the high volume of data, and the presence of noise and irrelevant information.

Natural language processing (NLP) techniques such as named entity recognition, part-of-speech tagging, and sentiment analysis have been widely used for event extraction from Twitter.

III. LITERATURE REVIEW

Several studies have explored the task of event extraction from Twitter. A study by Sakaki et al. (2010) proposed a method for detecting earthquakes from Twitter by analyzing the frequency and location of tweets. The study showed promising results in detecting earthquakes in real-time and could be used for early warning systems.

Another study by Ritter et al. (2012) proposed a method for identifying events and their attributes such as type, location, and time from Twitter. The study used a combination of rule-based and machine learning approaches to extract events and achieved an F1 score of 0.69.

Majumder et al. (2018) proposed an approach for extracting disaster-related events from Twitter using a combination of deep learning and rule-based techniques. The study used a convolutional neural network to extract relevant features from tweets and a rule-based approach to extract events. The proposed approach showed promising results in identifying disaster-related events.
In recent years, deep learning techniques such as recurrent neural networks (RNNs) and transformers have been used for event extraction from Twitter. Liu et al. (2020) proposed a transformer-based approach for extracting events and their attributes from Chinese social media platforms. The study achieved an F1 score of 0.65, outperforming traditional machine learning approaches.

In a recent study by Saha et al. (2021), a hybrid approach combining deep learning and rule-based methods was proposed for event extraction from Twitter. The study used a transformer-based model for encoding tweets and a rule-based approach for extracting events. The proposed approach outperformed traditional machine learning approaches and achieved an F1 score of 0.67 for event extraction.

Another recent study by Li et al. (2021) proposed a self-attention-based model for event extraction from social media. The study used a self-attention mechanism to capture the importance of each word in a tweet and a convolutional neural network to extract features from the tweet. The proposed model achieved state-of-the-art performance on the ACE-2005 dataset, a benchmark dataset for event extraction.

In a study by Du et al. (2021), a multi-task learning approach was proposed for event extraction and classification from social media. The study used a transformer-based model to extract features from tweets and a multi-task learning framework to jointly extract events and classify them into predefined event types. The proposed approach achieved competitive results on two benchmark datasets for event extraction and classification.

In another recent study by Wang et al. (2021), a graph-based approach was proposed for event extraction from Twitter. The study used a graph convolutional network to capture the dependencies between words in a tweet and a multi-task learning framework to jointly extract events and their attributes. The proposed approach achieved state-of-the-art results on two benchmark datasets for event extraction from Twitter.
Overall, recent studies show that deep learning techniques such as transformers, self-attention mechanisms, and graph convolutional networks can be effective for event extraction from social media platforms such as Twitter.

IV. METHODOLOGY

There are several approaches to event extraction from tweets, including rule-based, machine learning, and hybrid methods.

Some of the key techniques used in event extraction from tweets include named entity recognition, semantic role labelling, and temporal analysis. Named entity recognition involves identifying entities such as people, organizations, and locations mentioned in tweets.

Steps involved:

- Raw data Analysis
- Conversion of Raw data to Dataframe
- Defining functions for future uses
- Translating Tweets to the English language
- Data Pre-processing (Cleaning Tweets)
- Extracting Entities from the tweets with their frequency
- Data Visualization

V. RESULTS

Visualizing the Entities and their count using graphs for better insights. I created a new data frame with the entities and their frequency count in descending order. Put the file location and with help of the to_csv function, save it in the directory.

VI. CONCLUSION

Data Cleansing and Entities Extraction were a crucial part of the code. Removing unnecessary elements makes tweets ready for further processes, and This part of the process consumes most of the time. After that, Entities Extraction is a key point in the code. Finding the right entities defines extracting the right phrases from the tweets on which the core meaning of tweets relies on.

VII. REFERENCES

https://paperswithcode.com/task/twitter-event-detection

https://link.springer.com/chapter/10.1007/978-3-642-31178-9_32

https://www.cse.iitd.ac.in/~mausam/papers/kdd12.pdf

https://aclanthology.org/E17-1076.pdf