Extract and Summarize Tweets

Chaitanya Agarwal (2018114003)
Tanvi Kamble (2018114004)
Apurvi Mansinghka (2019201093)
Jyoti Gambhir (2019201032)

Project Report

Abstract

We attempt to create a model that extracts and summarizes tweets on about 11000 unique tweets. We use TwitterScraper for extraction followed by Page rank algorithm for summarization of tweets that mostly belong to the hashtags #blacklivesmatter. An attempt is also made to perform sentiment analysis on the extracted tweets using BERT HuggingFace Model. The results obtained are satisfactory and we also take care of the problems faced

1 Introduction

Twitter has become a mainstream platform for users to express their opinions on any topic and engage in debates. A large amount of valuable real-time information is posted on twitter. However this information is dispersed among millions of tweets containing varied sentiments, and opinions of the masses. User debates and interactions yield content based a specific topic which is referred as a 'trend'. Twitter trend extraction aims at clustering/grouping tweets with similar themes that are generated in a short period. To effectively utilize this content, it is essential to not only extract the information, but also to summarize the extracted data to move forward for the required analysis.
The most straightforward approach to solve this problem is to utilise 'hashtags'. It is a type of metadata tag used on social networks such as Twitter and Instagram and other micro-blogging services. It lets users apply dynamic, user-generated tagging that helps other users easily find messages with a specific theme or content. We attempt to get insights on any latest trend or social movements by extracting the tweets with relevant hashtags. The extracted tweets will then be put though a custom pipeline to summarize content and analyse varying sentiments over time.

1.1 Problem Statement

Given a set of hashtags, scrape twitter for tweets related to them and analyse (summarisation, sentiment analysis etc.) the scraped content.

2 Pipeline Design

Implementation details are as follows:

2.1 Scraping Tweets

Experiments were being carried out with Tweepy and TwitterScrapper. With Tweepy there were limitations like API only allows you to retrieve tweets upto 7 days and is limited to scraping 18,000 tweets per a 15 min window.TwitterScrapper overcomes this drawback by allowing user to extract tweets which is limited by internet speed/bandwidth and the number of instances of TwitterScrapper you are willing to start.
Tweets have been extracted on the basis of search, user timelines, ids, range of ids, user ids, users ids considering user timelines and places

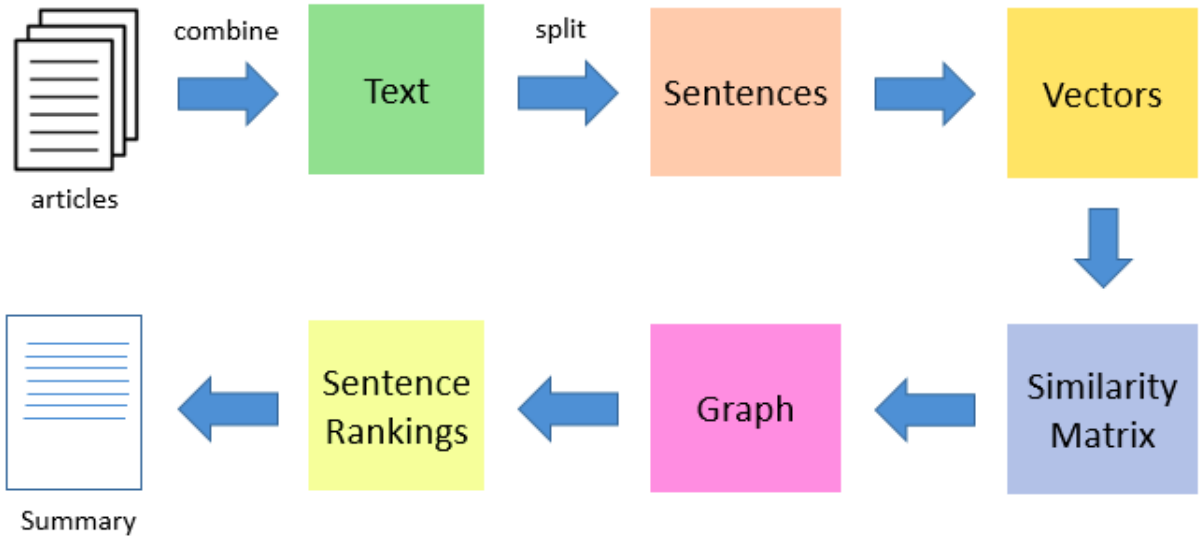
2.2 Pre-Processing

After the tweets are obtained, they're put through the pre-processing pipeline which involves replacement of abbreviations with their expansions, link and mentions removal, contraction replacement, hashtag removal, white-space and removal of punctuations. This helps us filter out elements which do not contribute to sentiment analysis and summarization.
After cleaning the text the text is tokenized and vectorized using GloVe Embeddings.
For the Sentiment Analysis, we also generate the token IDS using the BertTokenizer

2.3 Summarization of Tweets

We attempt to create event graphs where in the nodes are NE and event phrases and the weighted edges are the number of tweets in which the two nodes co-occurred in. We use this graph to summarize our tweets using the LexRank algorithm. The method used is presented in the Figure 1

Figure 1: Working of the LexRank Algorithm



2.4 Sentiment Analysis

Lastly we attempt to do Sentiment Analysis on the extracted tweets to find out the amount of positive and negative tweets and create a model that does so. We start off by using TextBlob, a pre trained model to find whether a tweet is positive or negative or neutral. Using these results we then train a BERT Transformer by HuggingFace to obtain a model.

3 Dataset

We extracted around 20,000 tweets corresponding the two trending Hastags #**blm** and #**blacklivesmatter**. Following are some statistics about the dataset after it was cleaned:

- Total Number of Tweets into consideration: **11,205**
- Total Number of duplicate Tweets : **11,077**
- Total Number of Words in the text are: **1,81,835**
- Number of Words in the longest Tweet: **60**
- Number of Words in the shortest Tweet: **0**
- Total Number of Unique Words in the text are: **26,190**
- Total Number of Hashtags in the text are: **5,707**
- Total Number of Mentions in the text are: **23,811**
- Total Number of Punctuation in the text are: **2,049**

The following was observed about the nature of the tweets after using the TextBlob:

Sentiment	**Percentage**
Positive	14.3
Negative	20.2
Neutral	65.5

4 Challenges

We expect pre-processing to be challenging due to a high amount of noise in the tweets. Moreover, event extraction, a notoriously tough NLU task, might yield upsetting results.

5 Results

Precision, recall, F1 score and support metrics are used for analysing the final results. Following are the observed values for sentiment classification :

	Precision	Recall	F1-Score	Support
Negative	0.94	0.92	0.93	156
Neutral	0.89	0.88	0.89	94
Positive	0.97	0.98	0.98	419
Accuracy			0.95	669
Macro Avg	0.93	0.93	0.93	669
Weighted Avg	0.95	0.95	0.95	669

Conclusion

To perform sentiment classification and analysis we require to extract large number of tweets, Tweeter-scapper is the most suitable scrapper for the above purpose. Page rank algorithm is used to summarize the tweets which uses backlinks and number of connects pages to find relevance of a a particular tweet. BERT model is used to prerform sentiment analysis on the summarized tweets.

References

[1] Goran Glavaš, Jan Šnajder, *Event graphs for information retrieval and multi-document summarization*, Expert Systems with Applications, Volume 41, Issue 15

[2] Wei Xu, Ralph Grishman, Adam Meyers, Alan Ritter, *A Preliminary Study of Tweet Summarization using Information Extraction*

[3] Zahra Majdabadi, Behnam Sabeti, Preni Golazizian, Seyed Arad Ashrafi Asli Omid Momenzadeh and Reza Fahmi, *Twitter Trend Extraction: A Graph-based Approach for Tweet and Hashtag Ranking, Utilizing No-Hashtag Tweets*, Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6213–6219 Marseille, 11–16 May 2020

[4] Roshni Chakraborty, Maitry Bhavsar, Sourav Dandapat, and Joydeep Chandra *A Network Based Stratification Approach for Summarizing relevant Comment Tweets of news articles*, Indian Institute of Technology, Patna