Apollo: Twitter Stream Analyzer of Trending Hashtags: A Case-study of #COVID-19

Mehwish Alam[1,2], Manuel Kaschura[2], and Harald Sack[1,2]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
[2] Karlsruhe Institute of Technology, Institute AIFB, Germany
{mehwish.alam,harald.sack}@fiz-karlsruhe.de,
manuel.kaschura@student.kit.edu

Abstract. This poster introduces a new tool named Apollo which analyzes textual information in the geo-tagged twitter streams of trending hashtags using sliding time window. It performs sentiment analysis as well as emotion detection of the opinions of the masses about a trending world wide topic such as #COVID-19, #ClimateChange, #BlackLivesMatter, etc. based on Knowledge Graphs. Apollo currently provides an interactive visualization of the analysis of the trending hashtag #COVID-19.

1 Introduction

Social Media has become one of the most effective ways for the populations to voice their opinions about ongoing major political, health, or any other kind of phenomena. Twitter is one of the popular social media channels where the users utilize hashtags for increasing the presence of their opinions about ongoing situations such as *#COVID-19*, *#BlackLivesMatter*, *#ClimateChange*, etc. The current global challenge faced by humanity is the COVID-19 pandemic. A huge amount of tweets are published each day targeting the related topics such as death rates, recovered cases, mental health issues, etc. These streaming twitter data can be used for many purposes such as monitoring mental health of the people during the lockdown, sentiments of the people suffering/recovering from the pandemic, agitation towards the measures taken by the governmental authorities for COVID-19 prevention, etc.

This paper particularly focuses on introducing a tool Apollo[3] which analyses the stream of English tweets related to *#COVID-19*. So far many other efforts have been put forth for analyzing the Twitter data. One of the early datasets [3] contains about 5 million tweets related to COVID-19, however, it only consists of the raw data. TweetKB [4] is another knowledge-base which performs sentiment analysis, entity linking, etc. of the tweets from October, 2019 to April, 2020 over

[3] http://covid-twitter-stream.fiz-karlsruhe.de/

8 million tweets. In [6], the authors perform sentiment classification using deep learning methods over tweets. However, none of the above reported studies have taken into account the stream of tweets. Due to huge amount of tweets being posted every minute all over the world, it is a challenge to process and analyze this information manually. This leads to the compulsion of introducing novel algorithms for automated analysis.

`Apollo` processes these twitter streams using sliding time window, retrieves geo-referenced tweets on the run and performs knowledge-aware sentiment analysis and emotion detection of the masses about the pandemic as well as frame detection over tweets using Framester [5].

2 Knowledge-Aware Twitter Stream Sentiment Analysis

Handling Twitter Streams. In order to handle twitter streams, time-based sliding window was used. A sliding window defines a time span with a fixed duration which goes back in time for this defined time span starting from current time. For instance, a sliding window of two minutes includes any events that have occurred in the past two minutes. This work uses Twitter Stream API [4] for obtaining twitter stream pertaining to particular hashtag such as #COVID-19 and variations of this hashtag such as "corona", "coronavirus", etc. The time span used for the current implementation is 30 minutes. The reason behind choosing 30 minute window is that it is extremely expensive to stream and process data over larger sliding time windows.

Pre-Processing Tweets. In the preprocessing step, hashtags were processed using word segmenter[5]. If the hashtag consists of multiple words, e.g., #blacklivesmatter, it was segmented into "black lives matter". The user mentions as well as URLs were removed because they are not used in the current analysis. For the retweets, the prefix "RT:" was removed. Emojis were mapped to the appropriate emotions manually. These mappings are available online[6].

Processing Geo-tagged Tweets. The locations of the tweets were converted to their co-ordinates using Geopy[7] which were later on used for plotting the emotions on the globe. For 30 minutes of streaming, more than 40,000 tweets were retrieved out of which 236 mentioned a place, 8 mentioned their co-ordinates, and 28002 contained location information. Only 0.03% of the tweets had their co-ordinates which led to the necessity of using external service such as Geopy. However, one of the bottle necks of Geopy is that it has limited number of requests per minute and for each request it takes 10-20 seconds which led to delays in the computation. In future, Apollo will use paid services for better performance. Another problem encountered was that most of the tweets did

[4] `http://docs.tweepy.org/en/latest/streaming_how_to.html`
[5] `https://pypi.org/project/wordsegment/`
[6] `https://github.com/ISE-FIZKarlsruhe/TwitterStreamAnalysis`
[7] `https://geopy.readthedocs.io/en/stable/`

not contain location information, or had locations such as planet earth, parallel universe, the emoticon of a globe, etc.

Processing Full Text. Full text of each of the tweets was processed by performing Word Sense Disambiguation (WSD). Several algorithms have been proposed so far but `Apollo` currently uses Lesk algorithm. Other algorithms such as UKB and Babelfy will be used for future releases. After obtaining the WordNet synsets, Framester was used for linking the tweets to the whole linked data resources (linguistic or otherwise) contained in Framester.

Framester is a large linguistic linked open data including about 30 million RDF triples acting as a hub between FrameNet, WordNet, VerbNet, BabelNet, Predicate Matrix, etc. It leverages this wealth of links to create an interoperable and homogeneous *predicate space* formally represented using frame semantics. Framester uses a mapping between WordNet, BabelNet, VerbNet and FrameNet at its core using detour based approach, expands it to other linguistic resources transitively, and represents all of this formally. It further links these resources to important external ontological and linked data resources such as DBpedia, YAGO, DOLCE-Zero, schema.org, NELL, etc. Further links to DeepKnowNet topic signatures, as well as SentiWordNet [2] and DepecheMood [7] mood mappings, are also available. Framester has been successfully used in downstream task such as knowledge reconciliation using frame embeddings [1].

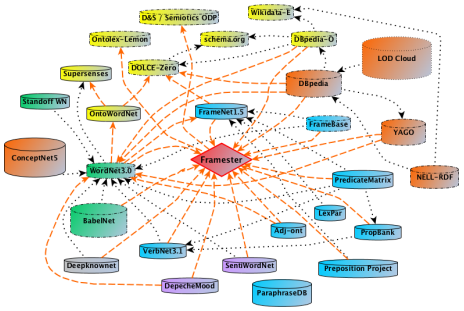
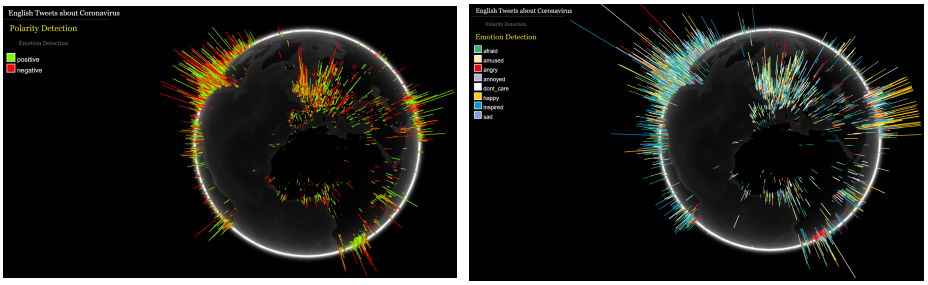


Fig. 1. Framester - Linguistic Linked Data Hub

Visualizing People's Sentiments and Emotions. The mappings contained in Framester about SentiWordNet were used to assign sentiment scores to the tweets. The overall score for the whole sentence was computed by averaging individual sentiment scores. In the presence of an emoji, a score of +0.5 or -0.5 was added to the numerator of the average and the denominator was increased by 1 (in case of one emoji). Similarly, for emotion detection the mappings between WordNet and DepecheMood (unique to Framester) were used. Finally, the aggregated score for each co-ordinate was visualized on 3-D globe which provides an interactive interface including zoom-in, zoom-out, etc. Figure 2 shows the final

visualization of the sentiments in `Apollo`. The legends on the left side describe the sentiments or emotions. The current version of `Apollo` only considers the English tweets which causes some parts of the globe to be darker meaning that there is only small amount of data to be analyzed due to higher frequencies of non-English tweets.



(a) Sentiments of masses about COVID-19 (b) Emotions of masses about COVID-19

Fig. 2. Interface of Apollo for visualizing sentiments (+ve and -ve) and emotions of the masses around the globe for a 30 minutes sliding window about `#COVID-19`

Word Frame Disambiguation. Currently, Apollo also enables the downloads which include the mappings of the tweets to FrameNet frames according to the TransX mappings between WordNet and FrameNet as provided by Framester (using `skos:closeMatch`). This further connects the tweets to other external resources in Framester such as DBpedia, DOLCE, etc. This process is referred to as Word Frame Disambiguation. Further mappings can be obtained by crawling Framester which is available through SPARQL endpoint[8].

3 Future Directions

In future, `Apollo` will be able to automatically extract trending hashtags. Different WSD algorithms such as UKB or Babelfy will be implemented. Twitter stream classification using machine learning algorithms for providing real-time analysis over larger time spans will also be implemented.

References

1. Alam, M., Recupero, D.R., Mongiovì, M., Gangemi, A., Ristoski, P.: Event-based knowledge reconciliation using frame embeddings and frame similarity. Knowl. Based Syst. **135**, 192–203 (2017). https://doi.org/10.1016/j.knosys.2017.08.014, `https://doi.org/10.1016/j.knosys.2017.08.014`

[8] `http://etna.istc.cnr.it/framester2/sparql`

2. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta. European Language Resources Association (2010)
3. Chen, E., Lerman, K., Ferrara, E.: Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. CoRR **abs/2003.07372** (2020), https://arxiv.org/abs/2003.07372
4. Dimitrov, D., Baran, E., Fafalios, P., Yu, R., Zhu, X., Zloch, M., Dietze, S.: Tweetscov19 - A knowledge base of semantically annotated tweets about the COVID-19 pandemic. CoRR **abs/2006.14492** (2020), https://arxiv.org/abs/2006.14492
5. Gangemi, A., Alam, M., Asprino, L., Presutti, V., Recupero, D.R.: Framester: A wide coverage linguistic linked data hub. In: Blomqvist, E., Ciancarini, P., Poggi, F., Vitali, F. (eds.) Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings. Lecture Notes in Computer Science, vol. 10024, pp. 239–254 (2016)
6. Li, I., Li, Y., Li, T., Álvarez-Napagao, S., García, D.: What are we depressed about when we talk about COVID19: mental health analysis on tweets using natural language processing. CoRR **abs/2004.10899** (2020), https://arxiv.org/abs/2004.10899
7. Staiano, J., Guerini, M.: Depeche mood: a lexicon for emotion analysis from crowd annotated news. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers. pp. 427–433. The Association for Computer Linguistics (2014). https://doi.org/10.3115/v1/p14-2070, https://doi.org/10.3115/v1/p14-2070