Proceedings of the 8th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'22)
Prague, Czech Republic - July 28- 30, 2022
Paper No. MHCI 114
DOI: 10.11159/mhci22.114

Social Media Event Detection Using SpaCy Named Entity Recognition and Spectral Embeddings

James A. von Albade and Joseph P. Salisbury
Riverside Research, Open Innovation Center
2640 Hibiscus Way, Beavercreek, Ohio, 45431, USA
jvonalbade@RiversideResearch.org; jsalisbury@riversideresearch.org

Abstract – Monitoring of public postings on social media platforms, such as Twitter, offers a valuable opportunity to detect events as they happen. Building off existing approaches for event detection, we evaluated the use of a pre-trained named entity recognition model followed by graph-based spectral clustering to detect events. We also examined transformer-based approaches for weighting the edges of the graph for event detection. Unlabeled events in the dataset we detected and verified are described.

Keywords**:** Twitter, event detection, microblogging, social media, manifold learning

1. Introduction

User-created social media content provides an invaluable resource for monitoring events across the globe. Using social media to detect events in a particular geographic area in real-time can enable more prompt responses to accidents, natural disasters, public health concerns, and security incidents. Twitter is an important social media platform for this purpose, given its ongoing popularity, volume of publicly accessible data, and concise format. As natural language processing (NLP) methods become more sophisticated, discovering more robust, nuanced, and scalable approaches to general event detection via Twitter continues to be an active field of research [1]–[6].

Here, we build off an existing approach, Metadata-Assisted Twitter Event Detection (MaTED) [1]. The MaTED approach consists of four main components, including detection of important phrases from tweets, creating temporal profiles of these phrases to identify "bursty" phrases, clustering bursty phrases with an aim to group related phrases about an event, and characterizing an event from the clusters obtained. We evaluated three major changes to this tweet-processing pipeline for event detection (**Fig. 1**).

First, we evaluate a simplified approach to named entity recognition (NER). MaTED introduced the use of DBPedia [7], [8] to identify named entities and WordNet [9] to assist in identifying event-specific words and phrases. To eliminate this requirement and provide a more lightweight alternative, we utilize a pre-trained spaCy [10] model for NER.

Secondly, we utilize an alternative similarity measure for clustering of bursty phrases. MaTED adopted an approach from Twevent [11], where the term frequency-inverse document frequency (TF-IDF) similarity metric is used for clustering bursty phrases using a graph-based clustering approach [12]. In contrast, we evaluate a semantic similarity metric based on Bidirectional Encoder Representations from Transformers (BERT) [13] to weight the edges of graphs.

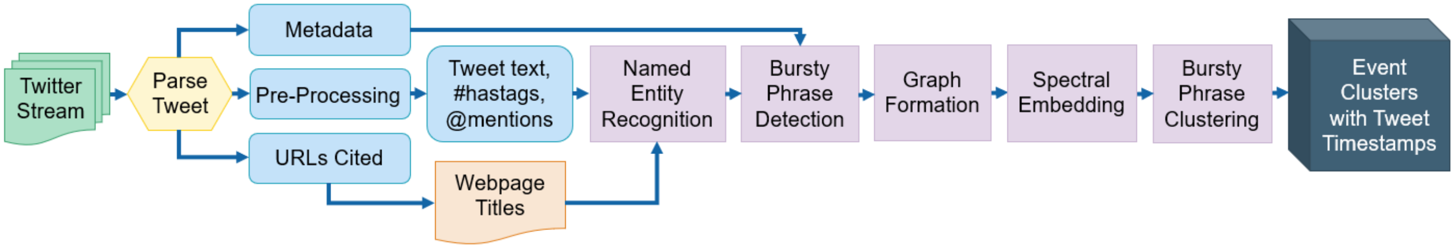

Fig. 1: Twitter event detection pipeline. Adapted from MaTED [1], except spaCy is used for named entity recognition, a transformer-based similarity measure is used during graph formation, and the graph is embedded in a lower-dimensional space before clustering.

Finally, we modify the graph-based clustering algorithm to utilize manifold learning. While MaTED simply forms a nearest neighbors graph and takes the connected components as events, we use spectral clustering that first embeds the graph in a lower dimensional space before clustering using a density-based algorithm like k-means. This renders the clustering more robust to the choice of the number of neighbors, while still allowing for automatic detection of the number of clusters.

2. Methods

To evaluate methods for Twitter event detection, we utilized the Events2012 corpus, which contains 120 million tweets and 506 labeled events [14]. We first divided tweets into time buckets. We chose one hour-long increments to identify shorter events. This choice may come at the expense of identifying events evolving over a longer timescale, but it may capture specific portions of these events. We then extracted tweet metadata and contextual information, such as favorites, retweets, and titles of linked webpages. We used spaCy's pretrained NER model [15], a transition-based parser, to extract key phrases. Each named entity was assigned a "burstiness" score within each time window corresponding to how often it appeared in that window relative to every other window [1]. This score is multiplied by the log of the number of followers, favorites, and retweets to weight tweets by engagement similar to MaTED's approach. A subset of the entities with the highest scores was selected as those potentially corresponding to events. As in MaTED, we selected $\sqrt{n}$, where n is the number of entities.

Entities in each window were associated with one another to identify events through a graph-based clustering algorithm. Each entity was identified as a node in a graph connected to a subset of other nodes by edges. We evaluated two different similarity metrics for edge formation. In the first, as in MaTED, for each entity we concatenated all tweets containing that entity into one document. We then calculated the TF-IDF similarity between each pair of entity documents, and each entity was connected by an edge to its k most similar entities. Alternatively, we evaluated a transformer-based similarity metric. We anticipated that this would more effectively capture semantic similarity between tweets, as words with similar meanings generally have similar embeddings in the pretrained model. A tweet T is fed to a BERT model pretrained on a large corpus of tweets [16], then the final hidden layer is averaged among all words to obtain a sentence embedding $E(T)$. The similarity between two documents D_1 and D_2 is then given by the average cosine similarity between tweets in the two documents:

$$S(D_1,\, D_2) = \frac{1}{|D_1||D_2|} \sum_{T_a \in D_1} \sum_{T_b \,\in\, D_2} \frac{E(T_a) \cdot E(T_b)}{\|T_a\| \, \|T_b\|} \tag{1}$$

In MaTED, clustering ends here – the events are identified by the connected components of the resulting graph. While this method is robust to non-convex clusters, this clustering is highly dependent on the choice of k. To remedy this, we utilized manifold learning [17]. We assume that while the graph is embedded in a high-dimensional space, it lies on a lower-dimensional manifold on which the clusters are approximately convex. To perform spectral clustering, we utilize the graph Laplacian:

$$L = D - A \tag{2}$$

where D is the degree matrix (in this case simply the identity) and A is the adjacency matrix. Since the adjacency matrix is not necessarily symmetric, we take the symmetric Laplacian:

$$L^{sym} = 0.5(L + L^T) \tag{3}$$

We then project onto the first n eigenvectors to give the embedding. This projection "unfolds" the manifold into the lower-dimensional space (Fig. 2). We then clustered using k-means with $k = 60$ for each time window. Since the number of events is not known, a more effective solution would be to use a clustering algorithm like OPTICS [18] or affinity propagation [19] that automatically identifies the number of clusters. However, k-means suffices for successfully identifying events that are not labeled in the dataset.

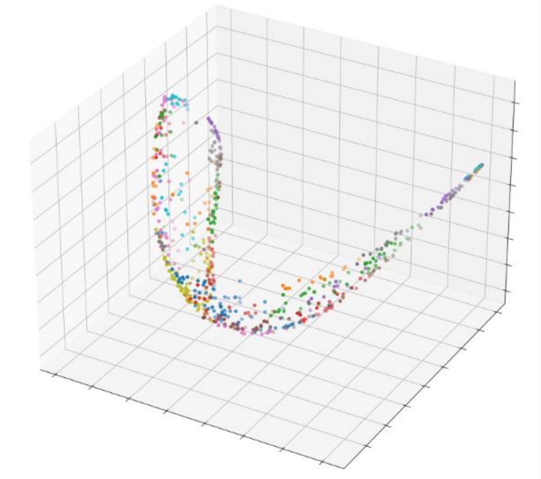

Fig. 2: 3D projection of entity nodes after spectral embedding, with colors corresponding to clusters. Note that because the embedding space is greater than three dimensions, clusters do not necessarily appear convex.

3. Results and Discussion

As the semi-supervised labeling of events in the dataset is not exhaustive, and in some cases is incorrect, precise quantification of model performance is intractable. However, we can see that many identified clustered in fact correspond to labeled events in the dataset. For example, when examining tweets from October 10th, 2012, one cluster of named entities identified by the system is “Pakistan, Malala, Taliban, Pakistani, Malala Yousafzai, Faridabad, Shyam Benegal, Radio Pakistan.” This successfully identifies the event “Malala Yousafzai, a 14-year-old activist for women's education rights is shot by Taliban gunmen in the Swat Valley.” Another cluster of words, “12in12, Chris Carpenter, Rockefeller Center, Washington, Washington D.C., 79 YEARS, D.C., Nationals, Carp, STLCards, Cards,” corresponds to the event “St. Louis Cardinals win their National League Divisional Series against the Washington Nationals.”

Our system additionally succeeds in identifying events that happened during the given timeframe but that are not in the dataset. For example, one cluster contains such entities as Robert Caro, Dave Eggers, and Louise Erdrich, all of whom were announced as nominees for the 2012 National Book Awards on October 10th [20]. While this is certainly an event that occurred during this time, it is not labeled in the dataset.

The BERT-based metric did not identify events that were not present in the TF-IDF metric. As calculating TF-IDF of documents is far faster, there is no obvious reason to prefer transformer embeddings. While these embeddings reflect semantic similarities between terms, this appears less relevant when events are typically characterized by proper nouns like names and places.

4. Conclusion

We described several modifications to the MaTED system that successfully identifies in an unsupervised manner many labeled and unlabled events in the Events2012 Twitter dataset. Our use of named entity recognition streamlines the phrase extraction in MaTED, while manifold learning allows for entity clustering to be better tailored to the geometry of events. The BERT-based metric did not improve performance compared to TF-IDF. While this pipeline looks at tweets only retroactively, it could easily be adapted for live event detection by identifying clusters of entities that are conspicuous in the past hour, for example, compared to previous time periods. Utilizing location metadata embedded in tweets would allow this system to identify events happening in a particular geographical area, providing insight into localized events as they unfold.

Acknowledgements

This research was supported by Independent Research and Development (IRAD) funds provided by Riverside Research.

References

[1] A. Pandya, M. Oussalah, P. Kostakos, and U. Fatima, “MaTED: Metadata-Assisted Twitter Event Detection System,” in *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, vol. 1237, M.-J. Lesot, S. Vieira, M. Z. Reformat, J. P. Carvalho, A. Wilbik, B. Bouchon-Meunier, and R. R. Yager, Eds. Cham: Springer International Publishing, 2020, pp. 402–414. doi: 10.1007/978-3-030-50146-4_30.

[2] I. Alsmadi and M. O’Brien, “Event Detection in Twitter: A Content and Time-Based Analysis,” *ArXiv211105274 Cs*, Oct. 2021, Accessed: May 09, 2022. [Online]. Available: http://arxiv.org/abs/2111.05274

[3] G. K K, J. Moni, J. G. Roy, A. C P, S. Harikrishnan, and G. G. Kumar, “Extreme Event Detection and Management using Twitter Data Analysis,” in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, Mar. 2022, pp. 917–921. doi: 10.1109/DASA54658.2022.9765076.

[4] Z. Mottaghinia, M.-R. Feizi-Derakhshi, L. Farzinvash, and P. Salehpour, “A review of approaches for topic detection in Twitter,” *J. Exp. Theor. Artif. Intell.*, vol. 33, no. 5, pp. 747–773, Sep. 2021, doi: 10.1080/0952813X.2020.1785019.

[5] R. Di Girolamo, C. Esposito, V. Moscato, and G. Sperlí, “Evolutionary game theoretical on-line event detection over tweet streams,” *Knowl.-Based Syst.*, vol. 211, p. 106563, Jan. 2021, doi: 10.1016/j.knosys.2020.106563.

[6] P. M. A. Y. Erfanian, B. R. Cami, and H. Hassanpour, “An evolutionary event detection model using the Matrix Decomposition Oriented Dirichlet Process,” *Expert Syst. Appl.*, vol. 189, p. 116086, Mar. 2022, doi: 10.1016/j.eswa.2021.116086.

[7] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, “DBpedia: A Nucleus for a Web of Open Data,” in *The Semantic Web*, Berlin, Heidelberg, 2007, pp. 722–735. doi: 10.1007/978-3-540-76298-0_52.

[8] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, “DBpedia spotlight: shedding light on the web of documents,” in *Proceedings of the 7th International Conference on Semantic Systems - I-Semantics ’11*, Graz, Austria, 2011, pp. 1–8. doi: 10.1145/2063518.2063519.

[9] C. Fellbaum, “Wordnet. the encyclopedia of applied linguistics.” John Wiley & Sons, Ltd, 2012.

[10] M. Honnibal and I. Montani, “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing,” *Appear*, vol. 7, no. 1, pp. 411–420, 2017.

[11] C. Li, A. Sun, and A. Datta, “Twevent: segment-based event detection from tweets,” in *Proceedings of the 21st ACM international conference on Information and knowledge management*, New York, NY, USA, Oct. 2012, pp. 155–164. doi: 10.1145/2396761.2396785.

[12] R. A. Jarvis and E. A. Patrick, “Clustering Using a Similarity Measure Based on Shared Near Neighbors,” *IEEE Trans. Comput.*, vol. C–22, no. 11, pp. 1025–1034, Nov. 1973, doi: 10.1109/T-C.1973.223640.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv, arXiv:1810.04805, May 2019. doi: 10.48550/arXiv.1810.04805.

[14] A. J. McMinn, Y. Moshfeghi, and J. M. Jose, “Building a large-scale corpus for evaluating event detection on twitter,” in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, New York, NY, USA, Oct. 2013, pp. 409–418. doi: 10.1145/2505515.2505695.

[15] “spaCy EntityRecognizer.” https://spacy.io/api/entityrecognizer (accessed May 20, 2022).

[16] D. Q. Nguyen, T. Vu, and A. T. Nguyen, “BERTweet: A pre-trained language model for English Tweets,” arXiv, arXiv:2005.10200, Oct. 2020. doi: 10.48550/arXiv.2005.10200.

[17] A. Ng, M. Jordan, and Y. Weiss, “On Spectral Clustering: Analysis and an algorithm,” in *Advances in Neural Information Processing Systems*, 2001, vol. 14. Accessed: May 20, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2001/hash/801272ee79cfde7fa5960571fee36b9b-Abstract.html

[18] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, “OPTICS: ordering points to identify the clustering structure,” *ACM SIGMOD Rec.*, vol. 28, no. 2, pp. 49–60, Jun. 1999, doi: 10.1145/304181.304187.

[19] B. J. Frey and D. Dueck, “Clustering by passing messages between data points,” *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[20] “2012 National Book Award Finalists Announced.” https://www.bookweb.org/news/2012-national-book-award-finalists-announced (accessed May 20, 2022).