CERIAS 2020
Annual Security Symposium ★ September 29 & 30
Purdue University, West Lafayette, Indiana

Twitter Topic and Bias Identification: A Graph-based Approach

PURDUE UNIVERSITY Discovery Park
Xiaonan Jing, Yi Zhang, Qingyuan Hu, Julia Rayz
Computer and Information Technology | AKRaNLU Lab

I. Abstract

An unusual fall semester at Purdue University began amid the COVID-19 pandemic last month with the new measures implemented to protect the boilermakers drew attention on social networks. Twitter serves as a data source for many Natural Language Processing (NLP) tasks. With a rapid changing online environment and a vast amount of textual data generated daily, it can be challenging to identify real-time topics on Twitter. Yet, tracing real-time topics is important for learning user interests and behaviors. Furthermore, potential content bias can be detected within the topics to help identify online security concerns. In this project, we are interested in detecting sub-topics on Twitter related to the popular "COVID-19" event at Purdue University. We employ graph structures, which are powerful tools for modeling the relationships between textual elements. A Graph-of-Words (GoW) based words association model is implemented to trace the daily Twitter content change through two weeks from Aug 19 to Sep 2. Particularly, we apply a Markov Clustering Algorithm (MCL) [1] along with a graph node removal approach to identify the daily content clusters. Additionally, we leverage FastText [2] word embeddings to identify content bias in tweets.

II. Data

- 23,942 tweets were fetched by keyword "Purdue" from Aug 19, 2020 to Sep 2, 2020

Date	Aug 19	Aug 20	Aug 21	Aug 22	Aug 23	Aug 24	Aug 25	Aug 26
Total	1,138	2,848	2,466	1,407	990	1,253	1,776	1,328
Ratio	3.43%	3.34%	4.14%	4.83%	2.83%	5.75%	3.43%	2.48%
Nodes	139	218	167	177	69	189	204	101
Date	Aug 27	Aug 28	Aug 29	Aug 30	Aug 31	Sep 1	Sep 2	
Total	1,028	1,454	786	616	1,191	2,468	3,193	
Ratio	3.21%	5.02%	3.44%	3.41%	4.53%	1.74%	0.60%	
Nodes	65	137	72	56	122	128	113	

- Total: Total number of tweets per day
- Ratio: number of tweets contain COVID-19 related hashtags per day / Total
- Nodes: Number of nodes in graph (NOUN and NAME ENTITY only)

III. Method

Topic Detection. Graph-of-Words (GoW) is a common word representation method inspired by the traditional Bag-of-Words (BoW). Typically, the vertices in a GoW represent the words from a corpus. In addition, the edges encode the co-occurrence association between the words in BoW. We adopt the GoW construction from [3] and generate a graph containing nouns and named entities from tweets for each day. The Markov Clustering algorithm (MCL) [2] partitions a graph by simulating probabilistic random walks in a graph to group highly connected nodes together. To remedy the drawback of MCL overclustering from a node which is highly connected with many other nodes, we first remove overly connected nodes.

Content Bias. Word2Vec family models are powerful in finding context based word associations. Words appearing in similar contexts will receive a higher cosine similarity score based on the distributional hypothesis that "a word is characterized by the company it keeps" [4] Following this hypothesis, user generated content bias can be identified from the data.

V. Results and Discussion

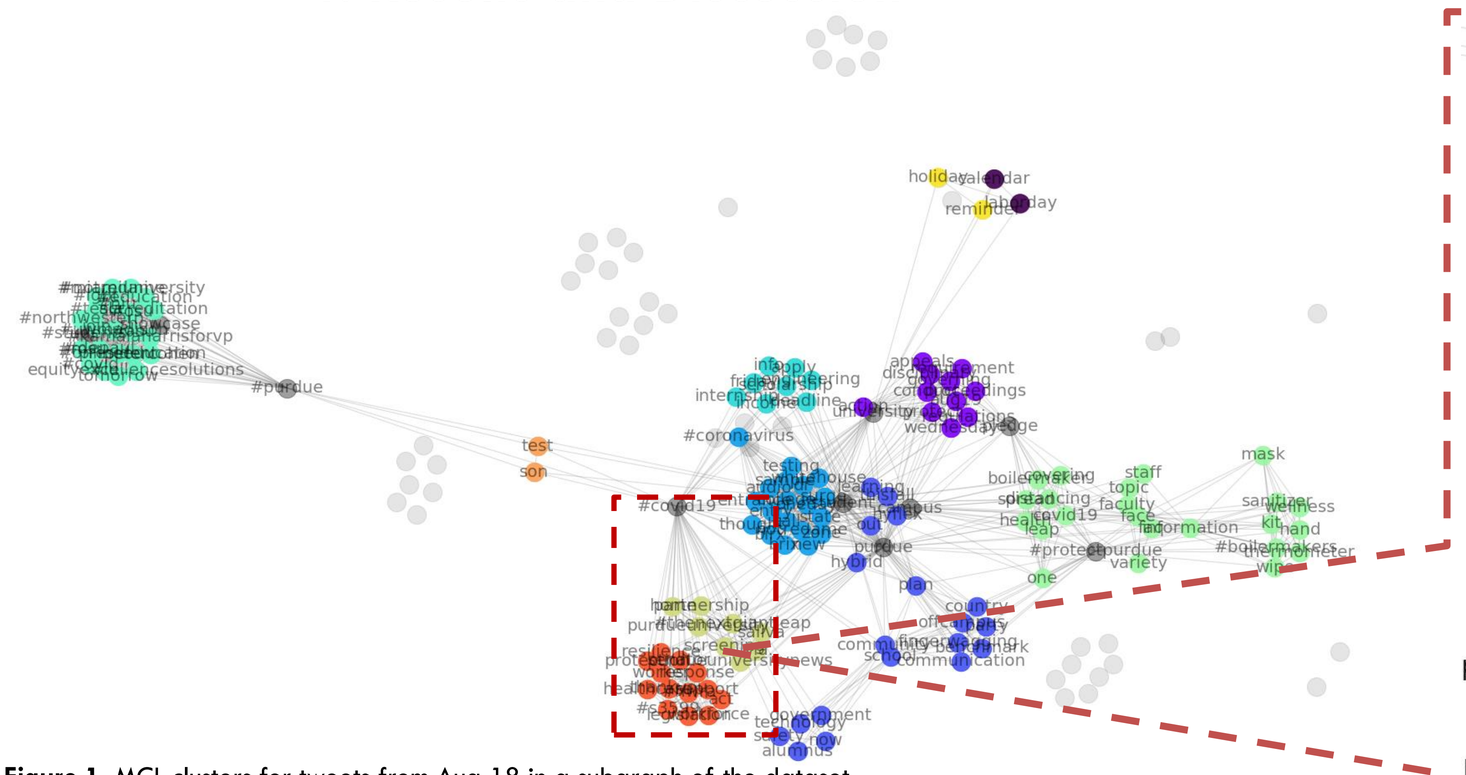

Figure 1. MCL clusters for tweets from Aug 18 in a subgraph of the dataset

Figure 2. The details of red and yellow clusters from **Figure 1**.

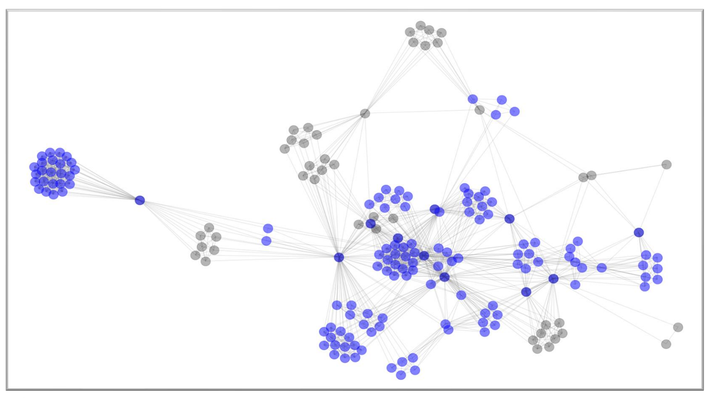

Figure 3. Blue nodes denote content of tweets for Aug 18

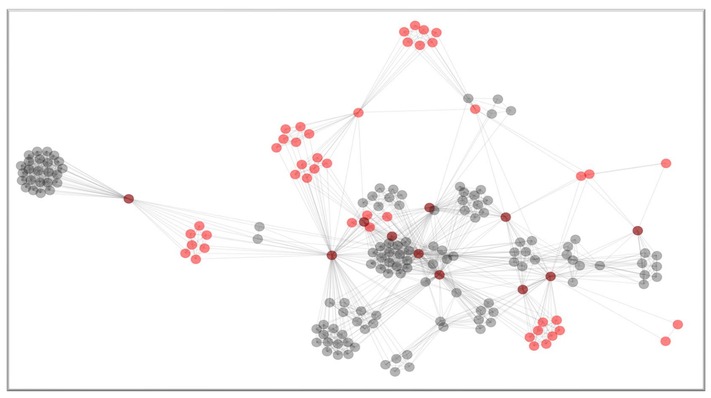

Figure 4. Red nodes denote content of tweets for Aug 30

Figures 1 and 2 demonstrate the GoW for COVID-19 related tweets collected for Aug 19th. Figure 2 zooms in on two clusters from Figure 1. Using the proposed approach, nodes that are highly correlated are clustered in different groups as illustrated by the colors. For instance, word "protection" and "healthcare" co-occurred in a tweet, and they are both colored in red. Words that are colored in gray, e.g. "#covid19", represent highly connected words that were removed before applying MCL. By the examining the words in Figure 2, it can be derived that the red cluster elaborates on "support/protection of workplace" and that the yellow cluster indicates topic on "COVID screening". On the other hand, Figure 3 and 4 provide an insight on how the contents of the tweets vary by date. It can be seen that the highlighted nodes on Aug 18 differ largely from that of Aug 30. By comparing the MCL clusters in different graphs, it is feasible to explore real-time topic changes in a given timeframe.

Content Bias. We determined some user generated content bias from our dataset by examining top similar words of several hashtags. For instance, "#skincare" is labeled as the second most similar word of "#pandemic" by the FastText model, which is incomprehensible in the context of COVID-19. However, after examining the data , it appears that "#pandemic" and "#skincare" co-occurred in the same tweets several times. An example is:

> **Her Campus Purdue** @HerCampusPurdue · Aug 31
> Check out Andi's article on how to treat your skin the right way! During a ~pandemic~ edition ✨#hercampus #purdue #hc #hercampuspurdue #article #onlinemagazine #magazine #pandemic #skincare

It should be noted that the bias may change based on one's default knowledge. By our default conception, "pandemic" should not be strongly associated with "skincare" as "pandemic" does not normally trigger an event related to "skincare"

VI. Future Work

- Currently, at different temporal time, our approach can only identify if there are changes to the temporal graph. Generating cluster labels can help us learn the topic changes through time better.
- GoW with only nouns and named entities cannot provide detailed information. Incorporating more part-of-speech can help enrich the information regarding a topic.
- In this project, we intentionally overfit the FastText model by training only on the total tweets collected in the time period. For word embeddings to be used in measuring the similarities within clusters, tuning a pre-trained model will serve as better means.

VII. References

[1] Van Dongen, S. M. (2000). Graph clustering by flow simulation (Doctoral dissertation).
[2] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.
[3] Meladianos, P., Nikolentzos, G., Rousseau, F., Stavrakas, Y., & Vazirgiannis, M. (2015). Degeneracy-based real-time sub-event detection in twitter stream. ICWSM, 15, 248-257.
[4] Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis.