In *Proc. Complex Networks & Their Applications XII,* Menton, France, Nov. 28-30, 2023, *Studies in Computational Intelligence,* pp. 215-224, vol. **1141**. Springer, Cham, Switzerland, Feb. 20, 2024

Analyzing Trendy Twitter Hashtags in the 2022 French Election

Aamir Mandviwalla[1,2], Lake Yin[1,2], and Boleslaw K. Szymanski[1,2]

[1] Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[2] Network Science and Technology Center, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

Abstract. Regressions trained to predict the future activity of social media users need rich features for accurate predictions. Many advanced models exist to generate such features; however, the time complexities of their computations are often prohibitive when they run on enormous data-sets. Some studies have shown that simple semantic network features can be rich enough to use for regressions without requiring complex computations. We propose a method for using semantic networks as user-level features for machine learning tasks. We conducted an experiment using a semantic network of 1037 Twitter hashtags from a corpus of 3.7 million tweets related to the 2022 French presidential election. A bipartite graph is formed where hashtags are nodes and weighted edges connect the hashtags reflecting the number of Twitter users that interacted with both hashtags. The graph is then transformed into a maximum-spanning tree with the most popular hashtag as its root node to construct a hierarchy amongst the hashtags. We then provide a vector feature for each user based on this tree. To validate the usefulness of our semantic feature we performed a regression experiment to predict the response rate of each user with six emotions like anger, enjoyment, or disgust. Our semantic feature performs well with the regression with most emotions having R^2 above 0.5. These results suggest that our semantic feature could be considered for use in further experiments predicting social media response on big data-sets.

Keywords: Computational social science · Social computing · Network science · 2022 French presidential election · Ukrainian war

1 Introduction

In recent years, social media data has been increasingly used to predict real-world outcomes. Data from platforms like Twitter, Reddit, and Facebook has

This preprint is a version of a publication of the same title presented at the *Complex Networks & Their Applications XII Conference*, Nov. 28-30, 2023, Menton, France. Wording may vary between versions. The copyrighted final version is accessible at `https://doi.org/10.1007/978-3-031-53468-3_18`

been shown to be valuable in predicting public sentiment or response towards many different topics. This information has been used across many different fields like predicting stock market price changes or movie popularity [3,14].

Social media platforms have continued to get more popular over time. Due to this, the size of social media datasets continues to increase. As the size of these datasets gets bigger and bigger, computational time complexity of the algorithms being used becomes a significant issue. Some of the most popular features used in social media predictions like sentiment analysis become prohibitively expensive when working with larger datasets [13,1].

In this paper we propose a method to generate features that can be used in social media predictions on big datasets. We create a weighted semantic network between Twitter hashtags from a corpus of 3.7 million tweets related to the 2022 French presidential election. A bipartite graph is formed where hashtags are nodes and weighted edges connect the hashtags reflecting the number of Twitter users that interacted with both hashtags. The graph is then transformed into a maximum-spanning tree with the most popular hashtag designated as its root node to construct a hierarchy amongst the hashtags. We then provide a vector feature for each user where the columns represent each of the 1037 hashtags in the filtered dataset and the value for each column is the normalized count of interactions for the user with that hashtag and any children of the hashtag in the tree.

To validate the usefulness of our semantic feature we performed a regression experiment to predict the response rate of each user with six emotions like anger, enjoyment, or disgust. The emotion data was manually annotated by a DARPA team created for the INCAS Program. We provide a baseline simple feature representing the counts the number of times a user interacts with each of the 1037 hashtags. Both the baseline and our semantic feature perform well with the regression with most emotions having R^2 above 0.5. The semantic feature outperforms the baseline feature on five out of six emotions with a p value of 0.05.

The rest of the paper is organized as follows. Section 2 details related works. In Section 3, we present the dataset used for experimentation. Then, Section 4 describes the methodology used in our paper. The design of experiments and their results are presented in Section 5, and the conclusions are discussed in Section 6.

2 Related Works

Analyzing Twitter using semantic networks has been done in the past with various methods to determine relationships between hashtags and their trends. For example, [17] considered two hashtags to be semantically related if an individual tweet contained both hashtags in the text. Similarly, [9] created a semantic network based on word co-occurrence within tweets. However, [16] presented an approach using a bipartite network between users and hashtags where an edge between a user node and a hashtag node was added if the user tweeted the hash-

tag at least once. This bipartite network was then projected into a monopartite network of hashtags. This approach is more applicable to our purposes because it captures the latent social network of the dataset. In addition, [16] focuses on a Twitter dataset taken from the 2018 Italian elections which is similar to our 2022 French election dataset.

Many studies have shown that semantic network features can be rich enough to use for regressions in a multitude of situations. In the field of psychology, semantic networks can be used to analyze a person's vocabulary to gain insight on cognitive states [7,4]. In terms of social media semantic networks, [10] used semantic networks generated from sentences as features for a time series regression to capture the volatility of the stock market. In general, these approaches involve creating a semantic network for each person or object in the study. The alternative approach is to create large-scale, singular semantic networks that can be used to describe all users. For example, [12] demonstrated a recommender system which used a large-scale word co-occurrence semantic network created from social media posts to recommend related social media posts to users. Such approach might be better for analyzing users since it can take advantage of the nuanced relationships between different social media communities, which cannot be done with an approach that only generates an individual semantic network for each user.

3 Data

We applied these enrichments to a dataset provided by DARPA INCAS program team that comprised 3.7 million French language tweets from 2022. This dataset was collected such that each tweet is relevant to the discussions that arose during the 2022 French presidential election. After pruning, this dataset contains 1037 hashtags and 389,187 users.

4 Methods

4.1 Semantic Network Generation

We performed several steps to prepare the Twitter data and create a semantic network.

Preprocessing The corpus of Tweets was first cleaned by removing URLs with regular expression and French stop words using the NLTK Python library [5]. Each Tweet was tokenized by converting all words to lowercase, removing digit-only words, and removing punctuation, except for hashtags. After extracting a set of hashtags and corresponding occurrence counts, any hashtags with an occurrence count below the mean were removed from the set to focus on trendy hashtags.

Bipartite Graph Generation Using this set of hashtags, a bipartite graph was constructed between users and the hashtags where an edge indicates an interaction between a user and a hashtag. Following the technique introduced in [16], we implemented the bipartite graph as an adjacency list where a set of interacted hashtags is stored with each user. A user is said to *interact* with a trendy hashtag if the user retweets, quotes retweets, comments under, or posts a tweet that contains the hashtag word with or without the hashtag symbol. We chose this relaxed approach because we consider situations such as "france" versus "#france" to be semantically identical. The resulting bipartite graph was projected along the hashtags as a weighted semantic network where each node represents a trendy Twitter hashtag, and the weighted edges represent the shared audience of users between two trendy hashtags.

Edge Pruning Next, the bipartite graph was then converted into a maximum spanning tree (MST) to only consider the most important links between trendy hashtags. We had conducted multiple experiments with and without edge pruning and concluded that some form of edge pruning is essential for removing noisy edges. We tested a flat cutoff approach for excluding edges with a weight below a set cutoff, and the MST approach, achieving the best and most robust results with the MST. All graph operations were performed with the NetworkX Python library [11]. A visualization of the resulting MST can be seen in Figure 1.

4.2 Semantic User Enrichment

Each user in the dataset $u \in U$ is assigned a set containing the trendy hashtags they had interacted with using the previously described interaction criteria. For each user set S_u, and for each trendy hashtag $t \in S_u$ where $S_u \subset V$ given semantic network/graph $G(V, E)$, each adjacency list corresponding to t is converted into an adjacency vector $\mathbf{a}_t = (a_{t1}, \ldots, a_{tn})$ where $n = |V|$, and

$$a_{ti} = \begin{cases} \frac{w(t,m(i))}{c(t)} & \text{if } e(t, m(i)) \in E \\ 1 & \text{if } m(i) = t \\ 0 & \text{otherwise} \end{cases}$$

where $w : E \to \mathbb{N}$ is the weight of the edge $e(u, v) \in E$, $m : [0, n] \to V$ that maps each index to a trendy hashtag, and $c : V \to \mathbb{N}$ maps each trendy hashtag to the number of users that have interacted with it. The set of vectors for each user is then summed element-wise and then normalized by dividing by the L^2 norm of the summed vector. The result for each user is a vector representing this user interests in a trendy hashtag and related trendy hashtags weighted by the latent social network.

4.3 Baseline User Enrichment

To judge the utility of our semantic network enrichment, we devised a simpler baseline enrichment for comparison. Each user $u \in U$ is assigned a vector $\mathbf{a}_u =$

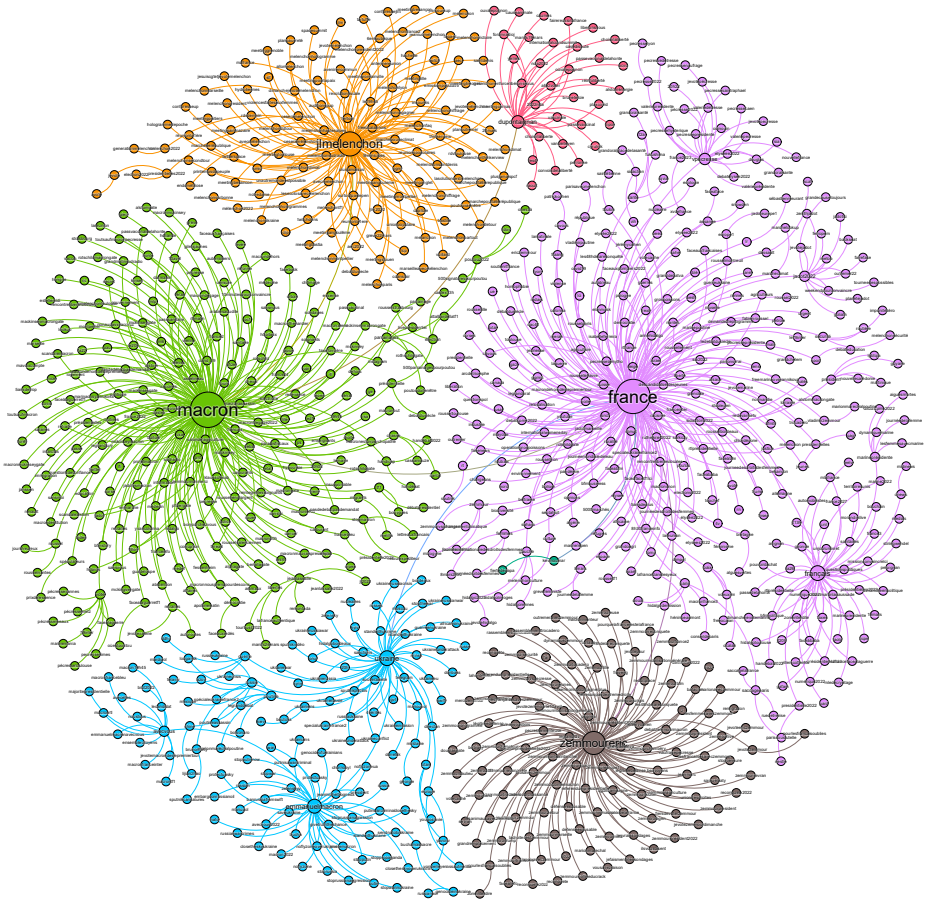

Fig. 1. Maximum-Spanning Tree of trendy hashtags in the 2022 French Presidential Election. Nodes represent trendy hashtags. Weighted edges between nodes represent the number of users that interacted with both trendy hashtags. Node size is based on weighted degree. Node color is based on modularity class after applying Louvain community detection and provided to help distinguish groups of similar nodes visually. From the root node #france, a hierarchy amongst trendy hashtags is formed, with clear distinction between different presidential candidates.

$(a_{u1}, \ldots, a_{un})$ where $n = |V|$, and $a_{ui} = k(u, m(i))$ where

$$k(u,t) = \begin{cases} 1 & \text{if } u \text{ interacts with trendy hashtag } t \\ 0 & \text{otherwise} \end{cases}$$

Each vector is normalized by dividing by the sum of the vector elements.

5 Results

5.1 Regression Experiment

To compare the enrichments, we conducted an experiment to test the performance of the enrichments in a regression task. We decided to test if a user enrichment could be used to predict the average "emotions" for each user. Each tweet in the dataset was annotated by the DARPA team with an array of six distinct emotions and an "other" value (representing fear, anger, enjoyment, sadness, disgust, surprise, and "none of the above" tag) where the sum of each array equals 1. For every user's tweets, we summed the emotion arrays, then divided the resulting array by its 1-norm, so that each array element follows $U(0, 1)$ and represents the probability of that user interacting with each emotion. Each user array was split into a set of emotion target variables, and each one was paired with the corresponding user enrichment method as the input variable. Only users with ≥ 10 tweets were included, resulting in 49,360 entries of input/target pairs for each emotion. Since this experiment is only meant to compare the different methods relative to each other with often minimal differences, we used the Scikit-learn implementation of linear regression [15].

5.2 Experiment Results

Emotion	Baseline R^2	Semantic R^2
Fear	0.222	0.229
Anger*	0.567	0.574
Enjoyment*	0.634	0.648
Sadness*	0.266	0.277
Disgust*	0.501	0.514
Surprise*	0.082	0.098
None	0.416	0.423

Table 1. Regression experiment results. (*: semantic > baseline; $p < 0.05$)

Overall, there is a clear pattern of improved performance when using the semantic enrichment instead of the baseline as seen in Table 1. All specific emotions, except fear and "none of the above" tag, statistically significantly improved

performance using the semantic method versus the baseline. Statistical significance was calculated using an F-test between the two models.

Interestingly, the baseline linear regression performed poorly on surprise emotion. The semantic enrichment performed significantly better, but still was weakly correlated with response. Intuitively, people would have various positive or negative views towards certain political trendy hashtags, which would correlate with most of the emotions. However, surprise cannot easily be categorized as on the positive or negative binary spectrum, which could explain why the linear regression performed poorly in those cases. Previous research on sentiment analysis has also shown notably lower performance when predicting surprise [6,18].

5.3 Semantic Analysis

To analyze which trendy hashtags are most associated with improvement with the semantic enrichment, we decided to filter for the top 10% of users that saw the most improvement in prediction accuracy between the baseline regression and the semantic regression. This was determined by the mean absolute error in emotion predictions versus the ground truth. Then we compared trendy hashtag occurrence rates for the top 10% users with the hashtag occurrence rates for the rest of the users. All hashtag occurrence rates were calculated based on direct interactions, like the baseline enrichment. We then selected the top 10 trendy hashtags that saw the largest increase in occurrence rate between the top 10% of users and the rest of the users.

Trendy hashtag	Bottom 90% rate	Top 10% rate
Paris	0.274	0.420
Youtube	0.328	0.470
Europe	0.458	0.600
Lci	0.227	0.348
Passe	0.250	0.370
Ukrainians	0.287	0.404
Jeunes	0.282	0.397
Liberté	0.371	0.485
Nucléaire	0.284	0.394
Immigration	0.248	0.357

Table 2. Trendy hashtags with the largest increase in occurrence rate between the top 10% of users and the rest of users. Occurrence rate is the proportion of users that directly interacted with that trendy hashtag. The top 10% of users is computed based on the improvement in prediction accuracy between the baseline regression and the semantic regression for those users.

Since the presence of these trendy hashtags in Table 2 result in more accurate emotion predictions with the semantic enrichment, this suggests that the users engaging with these trendy hashtags tend to engage with other emotionally

salient trendy hashtags, which would be more useful when predicting emotion levels. Given the severity of war, it would make sense that "Ukrainians" would be strongly connected to other highly emotionally salient trendy hashtags. A previous English language Twitter study about the Ukrainian war found that "Ukrainians" is a significant buzzword, so it is not surprising that the word reappears in French. In addition, that study identified the YouTube twitter account that was frequently mentioned in relation to the war, which would explain why it invokes strongly emotional trendy hashtags [19].

6 Conclusions and Future Work

We have connected semantic networks to the area of machine learning, demonstrating using a simple experiment that this can be used to consistently improve results on real-world data. To the best of our knowledge, this is the first time a semantic network feature for machine learning has been explored.

Future work can include specializing in such a framework to tackle specific problems. It is worth noting that the semantic network method used in this paper was designed with the constraints of the INCAS challenge in mind, which might not necessarily be the best way to utilize semantic networks when describing users in other situations. We are reporting these results simply to show that this method is a notable improvement over simpler approaches and it is worth investigating other applications in future work. One of the main drawbacks of this approach is the increased computational time associated with projecting the bipartite graph into a monopartite semantic network. However, there are distributed computing approaches to monopartite projection challenges, so this can be scaled for large scale applications involving many trendy hashtags [2].

During testing, we found that pruning small edges is important for removing noise from the final enrichments. We used the most aggressive pruning approach, by using a maximum spanning tree to only retain the strongest edges. This has the disadvantage of removing connections between different communities within the graph. It is possible that using a more sophisticated pruning approach could improve the quality of using semantic networks in this manner. Future work can take inspiration from knowledge graph edge pruning methods, which can account for domain information of the trendy hashtags [8].

Acknowledgements This work was partially supported by the DARPA INCAS Program under Agreement No. HR001121C0165 and by the NSF Grant No. BSE-2214216

References

1. Almuayqil, S.N., Humayun, M., Jhanjhi, N.Z., Almufareh, M.F., Khan, N.A.: Enhancing sentiment analysis via random majority under-sampling with reduced time complexity for classifying tweet reviews. Electronics **11**(21) (2022). https://doi.org/10.3390/electronics11213624

2. Asadi, M., Ghadiri, N., Nikbakht, M.A.: A scalable method for one-mode projection of bipartite networks based on hadoop platform. In: 2018 8th International Conference on Computer and Knowledge Engineering (ICCKE). pp. 237–242 (2018). https://doi.org/10.1109/ICCKE.2018.8566259
3. Asur, S., Huberman, B.A.: Predicting the future with social media. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. vol. 1, pp. 492–499 (2010). https://doi.org/10.1109/WI-IAT.2010.63
4. Beckage, N., Smith, L., Hills, T.: Semantic network connectivity is related to vocabulary growth rate in children. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 32 (2010)
5. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc." (2009)
6. Buechel, S., Hahn, U.: Emotion analysis as a regression problem–dimensional models and their implications on emotion representation and metrical evaluation. In: ECAI 2016, pp. 1114–1122. IOS Press (2016)
7. Chan, A.S., Salmon, D.P., Butters, N., Johnson, S.A.: Semantic network abnormality predicts rate of cognitive decline in patients with probable Alzheimer's disease. Journal of the International Neuropsychological Society **1**(3), 297–303 (1995). https://doi.org/10.1017/S1355617700000291
8. Faralli, S., Finocchi, I., Ponzetto, S.P., Velardi, P.: Efficient pruning of large knowledge graphs. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. p. 4055–4063. IJCAI'18, AAAI Press (2018)
9. Featherstone, J.D., Ruiz, J.B., Barnett, G.A., Millam, B.J.: Exploring childhood vaccination themes and public opinions on Twitter: A semantic network analysis. Telematics and Informatics **54**, 101474 (2020). https://doi.org/https://doi.org/10.1016/j.tele.2020.101474, https://www.sciencedirect.com/science/article/pii/S0736585320301337
10. Fronzetti Colladon, A., Grassi, S., Ravazzolo, F., Violante, F.: Forecasting financial markets with semantic network analysis in the COVID-19 crisis. Journal of Forecasting (2020)
11. Hagberg, A.A., Schult, D.A., Swart, P.J.: Exploring network structure, dynamics, and function using Networkx. In: Varoquaux, G., Vaught, T., Millman, J. (eds.) Proceedings of the 7th Python in Science Conference. pp. 11 – 15. Pasadena, CA USA (2008)
12. He, Y., Tan, J.: Study on SINA micro-blog personalized recommendation based on semantic network. Expert Systems with Applications **42**(10), 4797–4804 (2015). https://doi.org/https://doi.org/10.1016/j.eswa.2015.01.045, https://www.sciencedirect.com/science/article/pii/S0957417415000603
13. Kumari, S.: Impact of big data and social media on society. Global Journal for Research Analysis **5**, 437–438 (2016)
14. Pagolu, V.S., Challa, K.N.R., Panda, G., Majhi, B.: Sentiment analysis of Twitter data for predicting stock market movements. CoRR **abs/1610.09225** (2016)
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
16. Radicioni, T., Saracco, F., Pavan, E., Squartini, T.: Analysing Twitter semantic networks: the case of 2018 Italian elections. Scientific Reports **11**(1), 1–22 (2021)

17. Shi, W., Fu, H., Wang, P., Chen, C., Xiong, J.: #climatechange vs. #globalwarming: Characterizing two competing climate discourses on Twitter with Semantic Network and temporal analyses. International Journal of Environmental Research and Public Health **17**(3) (2020). https://doi.org/10.3390/ijerph17031062, https://www.mdpi.com/1660-4601/17/3/1062
18. Tsakalidis, A., Papadopoulos, S., Voskaki, R., Ioannidou, K., Boididou, C., Cristea, A.I., Liakata, M., Kompatsiaris, Y.: Building and evaluating resources for sentiment analysis in the Greek language. Language resources and evaluation **52**, 1021–1044 (2018)
19. Vyas, P., Vyas, G., Dhiman, G.: RUemo—The classification framework for Russia-Ukraine war-related societal emotions on Twitter through Machine Learning. Algorithms **16**(2) (2023). https://doi.org/10.3390/a16020069