Toxicity in Evolving Twitter Topics

Marcel Geller[1], Vítor V. Vasconcelos[2,3][0000−0003−3585−2417], and Flávio L. Pinheiro[1][0000−0002−0561−9641]

[1] NOVA IMS – Universidade Nova de Lisboa, Lisboa, Portugal
[2] Computational Science Lab, Informatics Institute, University of Amsterdam, Amsterdam, Netherlands
[3] Institute for Advanced Study, University of Amsterdam, Amsterdam, Netherlands

Abstract. Tracking the evolution of discussions on online social spaces is essential to assess populations' main tendencies and concerns worldwide. This paper investigates the relationship between topic evolution and speech toxicity on Twitter. We construct a Dynamic Topic Evolution Model (DyTEM) based on a corpus of collected tweets. To build DyTEM, we leverage a combination of traditional static Topic Modelling approaches and sentence embeddings using sBERT, a state-of-the-art sentence transformer. The DyTEM is represented as a directed graph. Then, we propose a hashtag-based method to validate the consistency of the DyTEM and provide guidance for the hyperparameter selection. Our study identifies five evolutionary steps or Topic Transition Types: Topic Stagnation, Topic Merge, Topic Split, Topic Disappearance, and Topic Emergence. We utilize a speech toxicity classification model to analyze toxicity dynamics in topic evolution, comparing the Topic Transition Types in terms of their toxicity. Our results reveal a positive correlation between the popularity of a topic and its toxicity, with no statistically significant difference in the presence of inflammatory speech among the different transition types. These findings, along with the methods introduced in this paper, have broader implications for understanding and monitoring the impact of topic evolution on the online discourse, which can potentially inform interventions and policy-making in addressing toxic behavior in digital communities.

Keywords: Social Media Platforms · Twitter · Topic Modelling · Topic Evolution · Discourse Toxicity

1 Introduction

The study of how topics in collections of documents evolve is not new [18, 8]. It follows naturally from the problem of automatically identifying topics [10] and then considering the time at which each document was produced and their lineage [2], i.e., when they emerge or collapse and their parent-child relationships. Such a description can provide insights into trends across many areas – such as in scientific literature [21, 27, 41], the web [6, 12], media [3, 26, 47], news [5, 49, 44, 36] – and the determinants of why some topic lineages extend longer while others fall short.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4

Online social networks – such as Twitter, Facebook, or Reddit – provide a valuable resource to study human behavior at large [40], constituting a rich source of observational data of individuals' actions and interactions over time. Text-based corpora from discussions on OSN can also be studied from a topic-level description and benefit from considering their temporal evolution. Contrary to collections of published documents – manuscripts or books – we often look into speech to better understand the intricacies of social dynamics and human behavior. The dynamics of topics emergence, merging, branching, persistence, or decline comes then as a consequence of our choices on which discussions we engage in and which not. In other words, our choices operate as a selective force that defines which topics prevail and which fade away from collective memory. Relevant in such dynamics are the language used within a topic, their efficiency in carrying information, and the resulting perception actors have of speech.

OSN have been used not only to revisit old theories but also document new phenomena such as social polarization and influence [16], information diffusion [42], the spread of disinformation [35] and information virality [22]. While OSNs provide access to large datasets, they come at the expense of requiring pre-processing and feature engineering to be studied [1, 15], of underlying biases that need to be accounted for [45], and experiments that need to be designed [43]. Many techniques have become popularly adopted to address such challenges: text-mining and machine learning methods have been used to estimate the Sentiment [34, 32, 37], Morality [24, 28, 4], or Toxicity [17] load in speech; network analysis [19] is often used to study patterns of information diffusion, connectivity, and community structure within Twitter.

Given this background, it is pertinent to ask how speech and associated features can modulate the evolutionary dynamics of topics over time. In this paper, we look at a large corpus of geolocated Tweets from New York (USA) to study the extent to which the evolution of topics is modulated by the toxicity of the embedded discourse. We use Topic Modelling methods and clustering techniques to track the emergence, branching, merging, persistence, and disappearance of topics from the social discussion. Specifically, we focus on discourse toxicity and its impact on topic evolution. Toxicity, in the context of online communication, refers to the presence of harmful, offensive, or aggressive language within a text. It encompasses a wide range of negative behaviors and expressions, such as hate speech, profanity, targeted harassment, personal attacks, threats, discriminatory language, and other forms of abusive or derogatory communication. Toxicity can manifest in various degrees, from mildly offensive remarks to extreme cases of online harassment and cyberbullying.

Our goal is to analyze if the Toxicity of topics tends to drift into higher/lower levels throughout their evolution and if there is any association between toxicity level and the topic's popularity. We contribute to better understand and effectively detect toxicity in online discourse, which is crucial for evaluating the health of digital communication ecosystems and informing policy-making and advancements in the domain of computational social science.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4

2 Related Work

OSNs have become a valuable source of observational data to study human behavior and social dynamics [31, 7]. Among various OSNs, Twitter, a microblogging online social network, has attracted significant attention from academic researchers. Its unique features include short posts limited to 280 characters, high frequency of posting, real-time accessibility to a global audience, and the provision of a free API allowing researchers to extract unfiltered and filtered content randomly or targeted from specific users or geolocations.

Although topic modeling is usually associated with the study and categorization of extensive collections of documents, it can also be used to identify sets of Tweets that share a common vocabulary. Naturally, dealing with short documents or tweets brings along several challenges, such as sparse co-occurrence of words across documents, informal language, high content variability, and noisy and irrelevant data. These challenges can affect the accuracy and reliability of topic models, so pre-processing techniques, such as text normalization, removal of stop words, and feature selection, are often applied to improve the quality of the input data[48].

Most studies approach Topic Modelling as a static process, neglecting the temporal dimension. However, when using Twitter data, we can track the evolution of discourse through the timestamp of each Tweet. In the particular case of topic evolution, several authors looked into dynamically modeling the public discourse. Malik et al. [33], propose a visual framework to study complex trends in public discussions over time. Abulaish and Fazil [2], studied topic evolution as an m-bipartite Graph and applied it to analyze the Evolution of tweets from Barack Obama, Donald Trump, and a Twitter Socialbot.

In both cases, the tweet corpus was divided into time bins, and Latent Dirichlet Allocation (LDA) Topic Modelling was utilized to compute topics of each time bin [9], which were then compared to subsequent time bins based on the cosine similarity of constructed topic embeddings. However, these approaches were tested on relatively small datasets of 16,199 and 3,200 Tweets, respectively, and a Topic Evolution Model has not yet been applied to studying the dynamics of speech toxicity.

In Alam et al [3], the temporal evolution of hashtag distributions for trending keywords is studied, the authors argue that combining hashtag information allows for a more suitable breakdown of topics underlying trending keywords while providing a context that offers better interpretability.

In addition to semantics, several speech characteristics can be inferred, which might be relevant in understanding both the intention of the writer but also how a reader perceives it. In that sense, perhaps the most popular text-based metric used is the sentiment of a text, which attempts to capture whether a person was expressing positive or negative thoughts through their speech [25]. However, the sentiment is somewhat narrow and lacks the nuance of the depth and variety of emotions. Other metrics that can be inferred from speech include the moral load of a document based on the Moral Foundations Theory or the embedded

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4

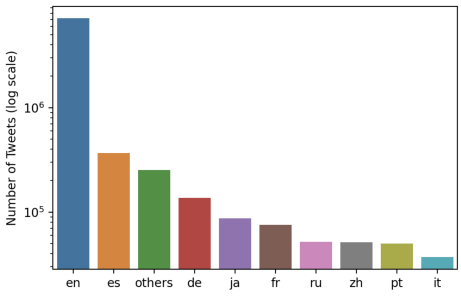

Fig. 1. Language of collected Tweets as classified by the fasttext language detector

toxicity of a text, which attempts to capture the existence of hateful or harmful language.

There are many proposed Text Mining approaches for detecting properties like toxicity or the moral load of a document. Most are based on a corpus of human-labeled documents[23]. Each document is transformed into a feature vector that embeds the relevant properties of the document. Then Machine Learning [39] is applied to train a Model on the feature vectors using the human-generated annotations as the ground truth.

The original contribution of this work lies in the construction of a dynamic topic evolution model to analyze the relationship between topic evolution and speech toxicity on geolocated tweets. By combining traditional static topic modeling approaches and sBERT sentence embeddings, we represent the topic evolution model as a directed graph and introduce a hashtag-based method for validation and hyperparameter selection. Our study expands on the existing literature by addressing the gap in understanding the dynamics of speech toxicity in topic evolution and its implications for online discourse and digital communities.

3 Data and Methods

We use a corpus of approximately 8 Million Tweets published between 2020/06/02 and 2020/11/03 and geo-located around the city of New York, USA. These cover a particularly interesting period of recent US history, which was marked by, for instance, the George Floyd protests, the strengthening of the Black Lives Matter movement and the 2020 presidential election race. The dataset was obtained using the free Twitter Academic API, with the only restriction being the geolocation of the tweets. Contrary to other studies, we have not focused on a particular topic or account. Instead, we resort to the collection in bulk of random tweets from the Twitter timeline.

Tweets are uniformly distributed in time, and on average we have 53,859 tweets per day (154 days in total). The dataset contains 364,918 unique hashtags with every hashtag occurring on average 4.81 times in the corpus. In order to further reduce the sparsity of the dataset and increase the number of token co-

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4

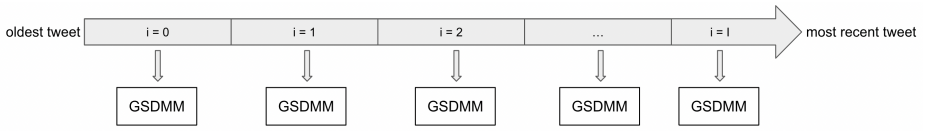

Fig. 2. The tweet corpus is divided into I time bins. The tweets in each bin are then clustered using the GSDMM Topic Modelling Algorithm.

occurrences between semantically similar documents, the following preprocessing steps were applied to the entire tweet corpus.

- Lowercasing of alphabetical characters.
- Removal of all emojis from the dataset.
- Deduplication of Tweets.
- Removal of all non-alphanumeric characters.
- Discarding of tokens containing less than 2 characters.
- Filtering out Tweets with less than 20 characters.
- Application of the fastText language detector [29, 30] to remove tweets not written in English.
- Lemmatization.

After applying these preprocessing steps the number of tweets in the corpus reduces by 37% to 5,197,172 Tweets.

Since the goal is to study topic evolution, we start by splitting the tweet corpus into non-overlapping time intervals of 10 days. Which leads to approximately 350.000 tweets per time interval, and eleven time intervals. The choice of non-overlapping time intervals over a sliding window method was done in order to reduce the number of time intervals to analyze and subsequently lower computational costs.

3.1 Topic Modelling and DAG Lineage

We use the Generalized Scalable Dirichlet Multinomial Mixture Model (GSDMM) to identify existing topics in the corpus [46]. We run GSDMM independently for each of the time bins, see Figure 2, and subsequently, we perform a topic linkage between the different time windows in order to obtain an approximation of topic evolution throughout the studied period, and as such their lineage.

The GSDMM is a variant of the popular LDA model, that is particularly useful for handling sparse data. GSDMM requires two hyperparameters, one that is used to represent the relative weight given to each cluster of words (α) and a second to represent the significance of each word in determining a document's topic distribution (β). Following Yin and Wang [46], we consider $\alpha = \beta = 0.1$ and we run GSDMM for 20 iterations for each time bin. GSDMM is able to automatically infer the number of topics, therefore it is simply necessary to initialize the number of topics, which we set to 120.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4

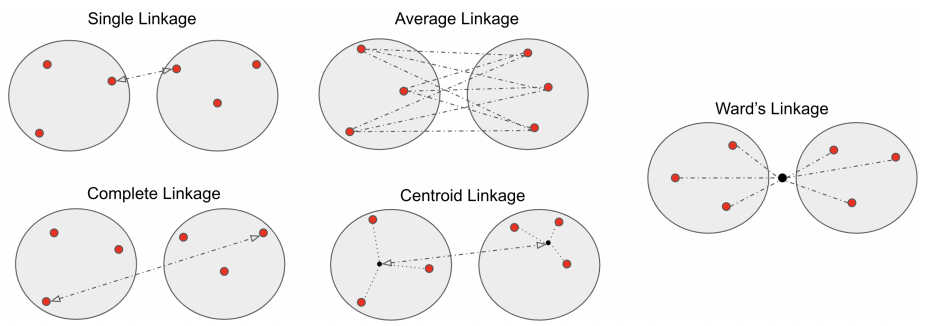

Fig. 3. Collection of Linkage Methods that can be used to quantify the distance between two Topics. Please note that it is required to have an embedding on the document level. In our case, we utilize sBERT to obtain a numerical vector for each Tweet.

At this stage, all time bins are considered independently from one another and topics are disconnected. Each tweet can be associated with a specific time bin and topic. To maintain consistent notations throughout the manuscript, let us define that M_j^i refers to a set of tweets that are associated with topic j at time i. Moving to the temporal evolution of Topics, we represent Topics and their parent-child relationship by means of an M-partite Directed Acyclic Graph (DAG). Each partition corresponds to all topics (j) of a given time bin (i), and edges connecting topics are an ordered pair of nodes (M_j^i, M_j^{i+1}) in two adjacent time windows.

In order to identify relationships between topics, we measure the semantic similarity between all topics from adjacent partitions. To that end, we use SentenceBERT (short sBERT), a pre-trained transformer-based model that can be used to encode each tweet into a high-dimensional feature vector[38]. The cosine distance between data points in this vector space is a metric to quantify the inverse semantic similarity of documents. A topic is a set of data points in the vector space that in order to compute the proximity of two topics needs to be compared to another set of data points in the same vector space created by sBERT. When computing the distance of two sets of data points we perform a Centroid Linkage approach [13]. We choose this approach over others due to its relative robustness against outliers when compared to, for instance, single or complete linkage methods and for being more computationally efficient than Ward linkage.

Finally, we use a threshold-based approach to select which candidate edges should be considered, in that sense we only consider edges that represent similarities (measured by the cosine similarity) between topics that are greater or equal to a threshold parameter ε.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4

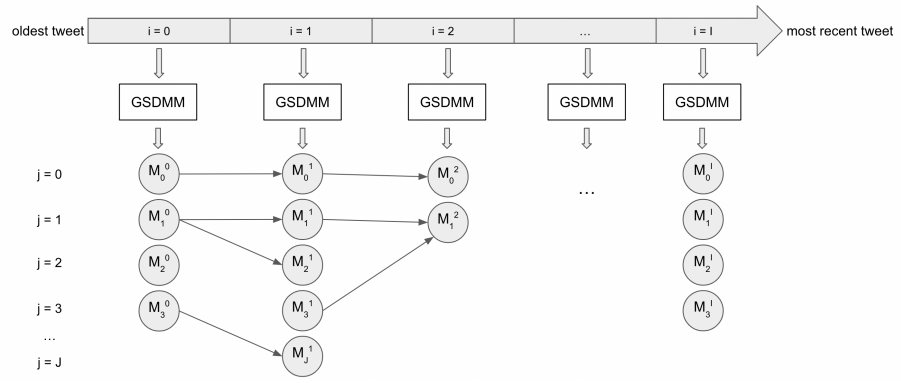

Fig. 4. Static Topic results are used to build a graph representation of the Topic Evolution. Topics of adjacent Time Bins are connected if a proximity threshold ε is exceeded.

3.2 Transition Types

By definition, a topic M_j^i can be connected to multiple topics from time bin $i+1$. Since multiple topics from time bin $i+1$ can have a similarity with M_j^i that exceeds ε, a parent topic node can have more than one child topic. This is also true the other way around. A child topic M_j^i can have more than one parent topic, i.e., if multiple topics from time bin $i+1$ are similar enough to M_j^i to exceed the proximity threshold ε. Hence, between time bin i and $i+1$ the following scenarios are possible to observe:

1. A Topic M_j^i splits into multiple Topics if the outdegree of M_j^i exceeds 1.
2. Multiple Topics merge into one Topic M_j^{i+1} if the indegree of M_j^{i+1} exceeds 1.
3. A Topic M_j^i stagnates if it has a single child topic M_j^{i+1} that has a single parent Topic. In other words, M_j^i needs to have an outdegree of 1 while its child Topic needs to have an indegree of 1.
4. M_j^i disappears if it has an outdegree of 0 (in other words: it has no child topic).
5. M_j^{i+1} emerges if it has an indegree of 0 (in other words: it has no parent topic).

These five scenarios are referred to as transition types throughout the rest of this paper. Moreover, by definition, topics in the first time bin do not emerge and topics in the last time bin do not disappear.

Figure 5 shows an illustrative example of the topic evolution. Taking topics at Bin 1 that are associated with the Black-Lives-Matter (BLM) movement, we follow their topic lineage for time 4 iterations. Each box corresponds to a topic and shows the five most popular hashtags. Topics that have BLM-related hashtags among the most popular hashtags are mostly connected to topics that

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4

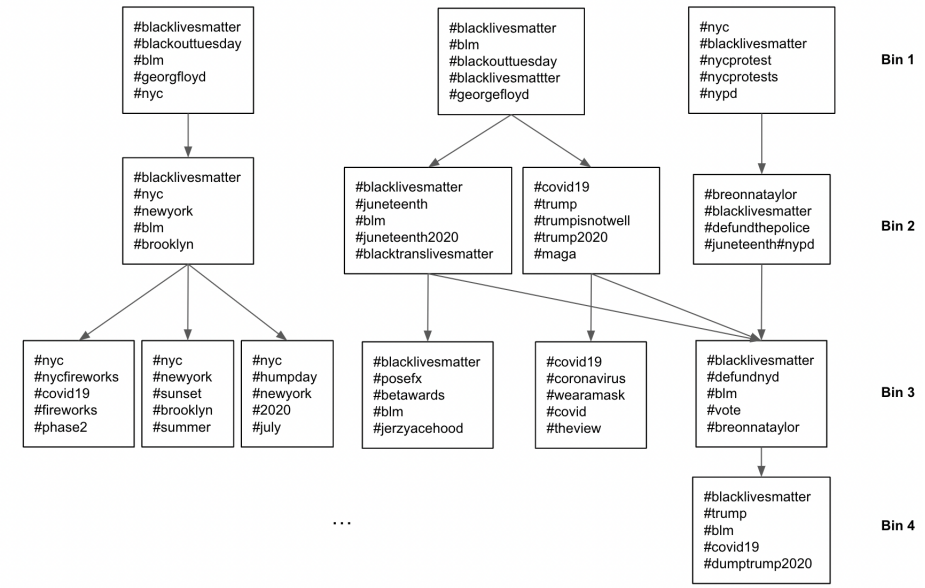

Fig. 5. Sample Topic Trajectories. Topics labeled with the 5 most frequent hashtags. A proximity threshold $\varepsilon = 0.8$ was chosen.

also contain BLM-related hashtags frequently. In this spot check a topic split was detected between Time Bin 1 and Time Bin 2, where BLM seems to split into two topics: one continues to be around the BLM events, and the second gears towards a more political topic involving Donald Trump.

As discussed above, the threshold ε is the only hyperparameter in the Topic Evolution approach presented in this paper. However, the selection of ε has a strong effect on the structure of the resulting DAG. To give an overview of the sensitivity of ε, figure 6 shows the proportion of each transition type in different values of ε. When selecting a value of ε close to its maximum value, only topics are connected that are almost identical between time bins. In this scenario, almost all transitions would be of type emergence and disappearance, because child or parent topics are due to the strict proximity requirement barely detected. In a scenario where ε is chosen to be a value close to its minimum value, almost all topics are connected to all topics from adjacent time bins. Therefore, in this case most topics are splitting and merging between time bins.

In order to validate the dynamic topic evolution DAG, we analyse whether the linkage done based on the BERT text embeddings is reflected by the distribution of hashtags across topics. We quantify the similarity of hashtags using the Jensen-Shannon-Divergence and compute it for each edge of the DAG. Starting from ε close to 1 only very similar topics are connected. The number of connections is therefore small while these few connected topics have a similar distribution of hashtags. As we decrease ε less similar topics are connected, which is why

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4

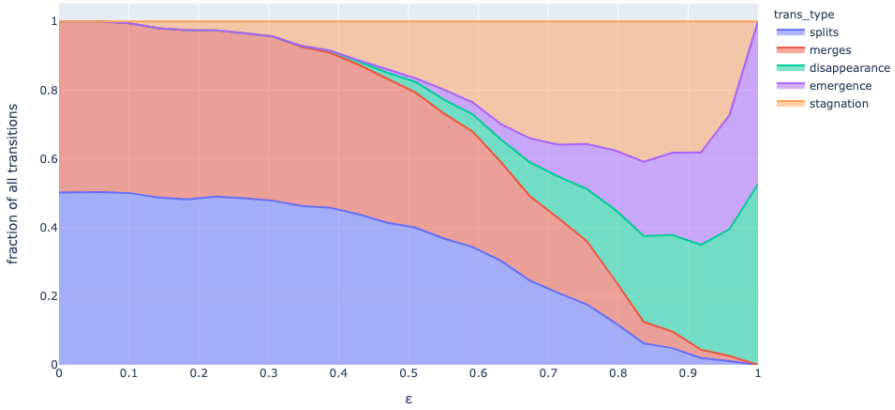

Fig. 6. Impact of the proximity threshold ε on the Transition Types Distribution

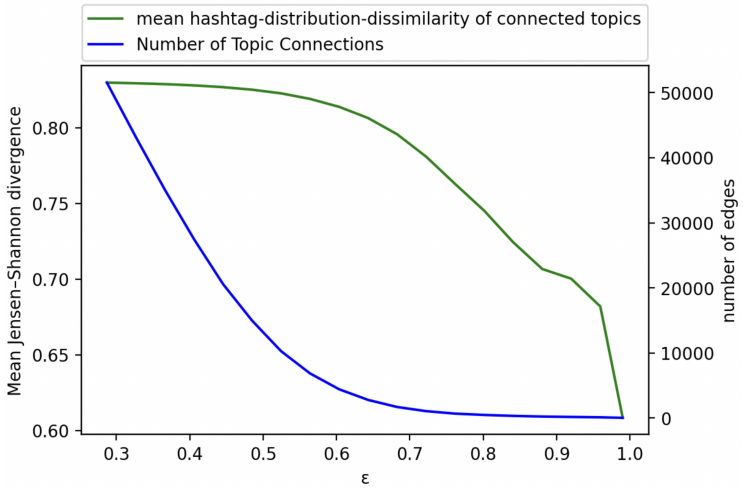

Fig. 7. Similarity of the distribution of hashtags across connected Topics in different value for ε

the number of edges in the DAG increases and the distribution of hashtags becomes less similar across connected topics. The fact that the hashtag similarity of connected topics is increasing in ε confirms the consistency of the model. Hashtags can arguably serve as a user-generated ground truth label for the Topic of a Tweet and in the given case support the linkage of topics discussed above.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4

4 Studying Toxicity in Topic Evolution

When analyzing the language used in a text corpus, toxicity means the existence of harmful or inflammatory language. Laura Hanu [20] proposed a machine learning model, Detoxify, that can estimate the level of embedded toxicity in a text. Although Detoxify is not trained on tweets but on a set of human-labeled Wikipedia comments, it is assumed that the model generalizes well enough to provide a good estimate for the level of toxicity in tweets. While Detoxify is a multi-label model that provides estimates for multiple speech characteristics, in the following analysis solely the toxicity of a document is used. While Detoxify returns a continuous level of toxicity between 0 and 1, the bimodal distribution suggests that it is reasonable to map the values to a binary variable. Binarizing the toxicity distribution is in line with the model's training on discrete human-annotated labels, which helps maintain consistency in the analysis and makes interpreting the results in relation to the original training data more straightforward. Using a cut-off threshold of 0.5, approximately 11% of the tweets are identified as toxic. This value is in line with other studies. Studies focusing on more controversial topics typically have a higher percentage of toxic tweets. Broad studies which focus on stronger definitions of toxicity, like hate speech and hateful content, have lower. For instance, a study by Davidson et al. (2017) [11] analyzed a dataset of over 25,000 tweets and found that around 5% of them contained hate speech. Another study by Founta et al. (2018) [14] analyzed a dataset of 80,000 tweets and found around 4% of abusive and hateful tweets in their random sample. Since we are focusing on the emergence of topics a higher fraction can help keep track of more nuanced topics.

The binarized toxicity can be included in the Topic Evolution DAG in the following manner. Each topic node is assigned with an attribute indicating the percentage of toxic tweets detected in the topic. Each edge of the Graph is assigned with an attribute called $\Delta toxicity$ that is calculated as follows, indicating the relative change in the percentage of toxic tweets.

$$\Delta Toxicity = \frac{Toxicity(childTopic)}{Toxicity(parentTopic)} - 1 \tag{1}$$

4.1 Toxicity per transition type

Each edge of the DAG belongs to at least one of the 5 transition types defined in chapter 3. Each merge and split transition, by definition, has multiple $\Delta Toxicity$ values assigned. Do topics tend to become more/less toxic when they merge/split/emerge/disappear/stagnate? The goal is to find out whether or not certain transition types have significantly different average $\Delta Toxicity$. The distribution of $\Delta Toxicity$ per transition type is presented in figure 8.

The results indicate no significant difference between the distributions of $\Delta Toxicity$ across transition types. The mean $\Delta Toxicity$ is near zero for all distributions, meaning that none of the transition types is correlated with a significant change in the percentage of inflammatory speech.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4

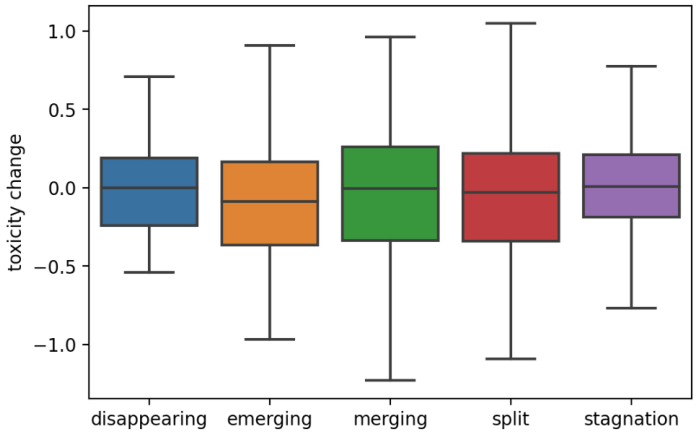

Fig. 8. Distribution of $\Delta Toxicity$ broken down by transition type. A proximity threshold of $\varepsilon = 0.8$ was chosen. Topics with a Popularity < 30 were discarded.

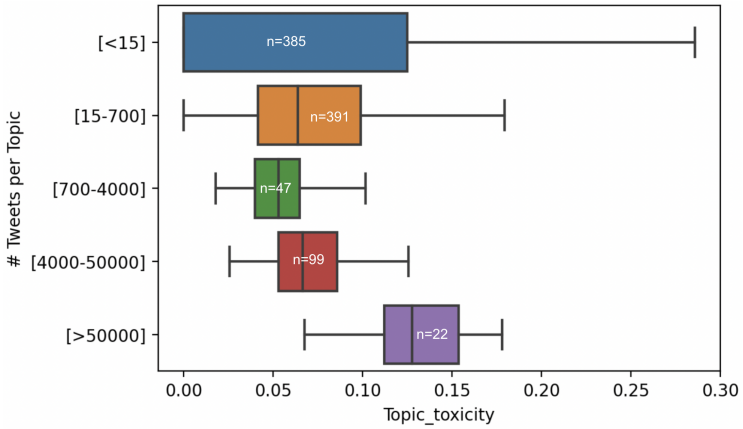

Fig. 9. Relationship between the Topic Toxicity and Topic Popularity (measured in Tweets per Topic)

4.2 Relationship Topic Popularity - Toxicity

We can measure topic popularity by assessing the number of tweets in each topic. We now analyse the relationship between Topic Popularity and Topic Toxicity. Do popular topics tend to be more or less toxic than niche Topics? As presented in Figure 10 and 9, our results indicate a parabola-shaped distribution of topics in the popularity-toxicity space. Ignoring micro topics with a popularity of less than 15 tweets, topic popularity is positively correlated with topic toxicity. Because of its right-skewed distribution, the topic popularity was log-transformed.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4

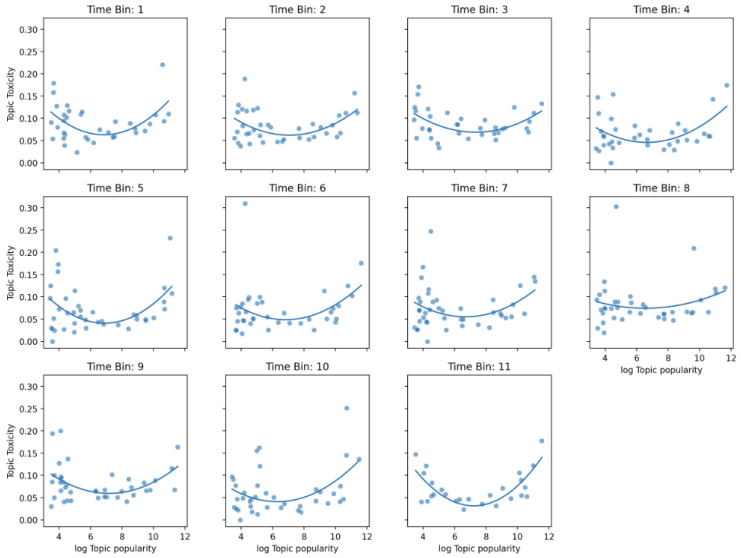

Fig. 10. Relationship between the Topic Toxicity and Topic Popularity grouped by time bin. Because of its skewed distribution the Topic Popularity was log transformed. Topics with an Popularity < 30 were discarded

5 Conclusion

In conclusion, this study introduces a dynamic Topic Evolution Modeling approach that represents topic trajectories in an M-partite-DAG. We analyzed the impact of the proximity threshold selection on the graph structure and validated the model's consistency using hashtags present in the tweets. Both manual spot checks of sample topic trajectories and quantification of hashtag similarity for all connected topics in the graph confirmed the model's consistency. Our findings reveal a positive correlation between topic popularity and toxicity, suggesting that viral topics tend to contain more inflammatory speech characteristics than niche topics. Moreover, we examined whether the level of speech toxicity evolves equally across all transition types. Our results indicate no significant difference in toxicity changes across the transition types, with changes being close to zero for all types. Despite the insights provided by our study, future research should aim to validate these results on other datasets with larger sample sizes and denser data. Additionally, while our current model focuses on topics emerging in adjacent time bins, it could be extended to consider the re-emergence of topics by comparing topics from non-adjacent time bins. This extension would remove the M-partite property of the graph but enable a more profound understanding of topic evolution.

In summary, our dynamic Topic Evolution Modeling approach offers a valuable tool for analyzing the relationship between topic dynamics and toxicity in Twitter discussions. This work contributes to the existing literature on online

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4

discourse and can inform future research on the nature of social media conversations and potential interventions to reduce toxicity and promote healthier digital spaces.

6 Acknowledgments

FLP acknowledges the financial support of the Portuguese Foundation for Science and Technology ("Fundação para a Ciência e a Tecnologia") through grant DSAIPA/DS/0116/2019.

References

1. Abidin, D.Z., Nurmaini, S., Malik, R.F., Rasywir, E., Pratama, Y., et al.: A model of preprocessing for social media data extraction. In: 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS). pp. 67–72. IEEE (2019)
2. Abulaish, M., Fazil, M.: Modeling topic evolution in twitter: An embedding-based approach. IEEE Access **6**, 64847–64857 (2018)
3. Alam, M.H., Ryu, W.J., Lee, S.: Hashtag-based topic evolution in social media. World Wide Web **20**, 1527–1549 (2017)
4. Araque, O., Gatti, L., Kalimeri, K.: Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. Knowledge-based systems **191**, 105184 (2020)
5. Bai, Y., Jia, S., Chen, L.: Topic evolution analysis of covid-19 news articles. In: Journal of Physics: Conference Series. vol. 1601, p. 052009. IOP Publishing (2020)
6. Bar-Ilan, J., Peritz, B.C.: A method for measuring the evolution of a topic on the web: The case of "informetrics". Journal of the American Society for Information Science and Technology **60**(9), 1730–1740 (2009)
7. Bello-Orgaz, G., Jung, J.J., Camacho, D.: Social big data: Recent achievements and new challenges. Information Fusion **28**, 45–59 (2016)
8. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning. pp. 113–120 (2006)
9. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
10. Boyd-Graber, J., Hu, Y., Mimno, D., et al.: Applications of topic models. Foundations and Trends® in Information Retrieval **11**(2-3), 143–296 (2017)
11. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the international AAAI conference on web and social media. vol. 11, pp. 512–515 (2017)
12. Derntl, M., Günnemann, N., Tillmann, A., Klamma, R., Jarke, M.: Building and exploring dynamic topic models on the web. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. pp. 2012–2014 (2014)
13. El-Hamdouchi, A., Willett, P.: Comparison of hierarchic agglomerative clustering methods for document retrieval. The Computer Journal **32**(3), 220–227 (1989)
14. Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., Kourtellis, N.: Large scale crowdsourcing and characterization of twitter abusive behavior. In: Proceedings of the international AAAI conference on web and social media. vol. 12 (2018)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4

15. Gani, R., Chalaguine, L.: Feature engineering vs bert on twitter data. arXiv preprint arXiv:2210.16168 (2022)
16. Garimella, V.R.K., Weber, I.: A long-term analysis of polarization on twitter. In: Eleventh international AAAI conference on web and social media (2017)
17. Georgakopoulos, S.V., Tasoulis, S.K., Vrahatis, A.G., Plagianakos, V.P.: Convolutional neural networks for toxic comment classification. In: Proceedings of the 10th hellenic conference on artificial intelligence. pp. 1–6 (2018)
18. Gohr, A., Hinneburg, A., Schult, R., Spiliopoulou, M.: Topic evolution in a stream of documents. In: Proceedings of the 2009 SIAM international conference on data mining. pp. 859–870. SIAM (2009)
19. Grandjean, M.: A social network analysis of twitter: Mapping the digital humanities community. Cogent Arts & Humanities **3**(1), 1171458 (2016)
20. Hanu, L., Unitary team: Detoxify. Github. https://github.com/unitaryai/detoxify (2020)
21. He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., Giles, L.: Detecting topic evolution in scientific literature: how can citations help? In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 957–966 (2009)
22. HOANG, T.A., LIM, E.P., ACHANANUPARP, P., JIANG, J., ZHU, F.: On modeling virality of twitter content.(2011). In: Digital Libraries: 13th International Conference on Asia-Pacific Digital Libraries, ICADL. pp. 24–27 (2011)
23. Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A.M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., et al.: Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. Social Psychological and Personality Science **11**(8), 1057–1071 (2020)
24. Hopp, F.R., Fisher, J.T., Cornell, D., Huskey, R., Weber, R.: The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. Behavior research methods **53**, 232–246 (2021)
25. Hu, R., Rui, L., Zeng, P., Chen, L., Fan, X.: Text sentiment analysis: A review. In: 2018 IEEE 4th International Conference on Computer and Communications (ICCC). pp. 2283–2288. IEEE (2018)
26. Hu, Y., Xu, X., Li, L.: Analyzing topic-sentiment and topic evolution over time from social media. In: Knowledge Science, Engineering and Management: 9th International Conference, KSEM 2016, Passau, Germany, October 5-7, 2016, Proceedings 9. pp. 97–109. Springer (2016)
27. Jo, Y., Hopcroft, J.E., Lagoze, C.: The web of topics: discovering the topology of topic evolution in a corpus. In: Proceedings of the 20th international conference on World wide web. pp. 257–266 (2011)
28. Johnson, K., Goldwasser, D.: Classification of moral foundations in microblog political discourse. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers). pp. 720–730 (2018)
29. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 (2016)
30. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
31. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al.: Computational social science. Science **323**(5915), 721–723 (2009)
32. Li, N., Wu, D.D.: Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision support systems **48**(2), 354–368 (2010)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4

33. Malik, S., Smith, A., Hawes, T., Papadatos, P., Li, J., Dunne, C., Shneiderman, B.: Topicflow: Visualizing topic alignment of twitter data over time. In: Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining. pp. 720–726 (2013)
34. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal **5**(4), 1093–1113 (2014)
35. Murayama, T., Wakamiya, S., Aramaki, E., Kobayashi, R.: Modeling the spread of fake news on twitter. Plos one **16**(4), e0250419 (2021)
36. Neo, S.Y., Ran, Y., Goh, H.K., Zheng, Y., Chua, T.S., Li, J.: The use of topic evolution to help users browse and find answers in news video corpus. In: Proceedings of the 15th ACM international conference on Multimedia. pp. 198–207 (2007)
37. Redhu, S., Srivastava, S., Bansal, B., Gupta, G.: Sentiment analysis using text mining: a review. International Journal on Data Science and Technology **4**(2), 49–53 (2018)
38. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
39. Roy, S., Pacheco, M.L., Goldwasser, D.: Identifying morality frames in political tweets using relational learning. arXiv preprint arXiv:2109.04535 (2021)
40. Salganik, M.J.: Bit by bit: Social research in the digital age. Princeton University Press (2019)
41. Song, M., Heo, G.E., Kim, S.Y.: Analyzing topic evolution in bioinformatics: investigation of dynamics of the field with conference data in dblp. Scientometrics **101**, 397–428 (2014)
42. Stai, E., Milaiou, E., Karyotis, V., Papavassiliou, S.: Temporal dynamics of information diffusion in twitter: Modeling and experimentation. IEEE Transactions on Computational Social Systems **5**(1), 256–264 (2018)
43. Tan, C., Lee, L., Pang, B.: The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. arXiv preprint arXiv:1405.1438 (2014)
44. Viermetz, M., Skubacz, M., Ziegler, C.N., Seipel, D.: Tracking topic evolution in news environments. In: 2008 10th IEEE Conference on E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services. pp. 215–220. IEEE (2008)
45. Yang, K.C., Hui, P.M., Menczer, F.: How twitter data sampling biases us voter behavior characterizations. PeerJ Computer Science **8**, e1025 (2022)
46. Yin, J., Wang, J.: A dirichlet multinomial mixture model-based approach for short text clustering. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (2014)
47. Zhang, Y., Mao, W., Lin, J.: Modeling topic evolution in social media short texts. In: 2017 IEEE International Conference on Big Knowledge (ICBK). pp. 315–319. IEEE (2017)
48. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings 33. pp. 338–349. Springer (2011)
49. Zhou, H., Yu, H., Hu, R., Hu, J.: A survey on trends of cross-media topic evolution map. Knowledge-Based Systems **124**, 164–175 (2017)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_4