MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Toxicity in Evolving Twitter Topics

Employing a novel Dynamic Topic Evolution Model (DyTEM) on Twitter data

Marcel Geller

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

TOXICITY IN EVOLVING TWITTER TOPICS

by

Marcel Geller

Dissertation presented as partial requirement for obtaining the Master’s degree in Advanced Analytics, with a Specialization in Data Science

Supervisor: prof Doutor Flávio Luis Portas Pinheiro

07 - 2023

PRIMARY REFERENCE

This thesis is largely based on the paper "Toxicity in Evolving Twitter Topics" by Marcel Geller, Flávio L. Pinheiro, and Vítor V. Vasconcelos. The paper was presented at the 23rd International Conference on Computational Science (ICCS 2023) in Prague, Czech Republic. It is included in the Springer's Lecture Notes in Computer Science series (Vol. 10476), as Chapter 4 (pp. 40-54) of the conference proceedings.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Marcel Geller

Zurich, 02.07.2023

ABSTRACT

This thesis presents an extensive investigation into the evolution of topics and their association with speech toxicity on Twitter, based on a large corpus of tweets, providing crucial insights for monitoring online discourse and potentially informing interventions to combat toxic behavior in digital communities. A Dynamic Topic Evolution Model (DyTEM) is introduced, constructed by combining static Topic Modelling techniques and sentence embeddings through the state-of-the-art sentence transformer, sBERT. The DyTEM, tested and validated on a substantial sample of tweets, is represented as a directed graph, encapsulating the inherent dynamism of Twitter discussions. For validating the consistency of DyTEM and providing guidance for hyperparameter selection, a novel, hashtag-based validation method is proposed. The analysis identifies and scrutinizes five distinct Topic Transition Types: Topic Stagnation, Topic Merge, Topic Split, Topic Disappearance, and Topic Emergence. A speech toxicity classification model is employed to delve into the toxicity dynamics within topic evolution. A standout finding of this study is the positive correlation between topic popularity and its toxicity, implying that trending or viral topics tend to contain more inflammatory speech. This insight, along with the methodologies introduced in this study, contributes significantly to the broader understanding of digital discourse dynamics and could guide future strategies aimed at fostering healthier and more constructive online spaces.

KEYWORDS

Twitter Research; Topic Modelling; Topic Evolution; Discourse Toxicity; Dynamic Topic Modelling

Sustainable Development Goals (SGD):

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

DAG	Directed Acyclic Graph
GSDMM	Gibbs Sampling Dirichlet Mixture Model
OSNs	Online Social Networks
SBERT	Sentence Bidirectional Encoder Representations Transformers
BLM	Black Lives Matter
LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
RoBERTa	Robustly Optimized BERT Approach
MultiNLI	Multi Genre Natural Language Inference
SNLI	Stanford Natural Language Inference
DyTEM	Dynamic Topic Evolution Model

1. INTRODUCTION

The study of how topics in collections of documents evolve is not new (Blei & Lafferty, 2006; Gohr et al., 2009). It follows naturally from the problem of automatically identifying topics (Boyd-Graber et al., 2017) and then considering the time at which each document was produced and their lineage (Abulaish & Fazil, 2018), 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 (He et al., 2009; Jo et al., 2011; Song et al., 2014), the web (Bar-Ilan & Peritz, 2009; Derntl et al., 2014), media (Alam et al., 2017; Hu et al., 2016; Zhang et al., 2017), news (Bai et al., 2020; Neo et al., 2007; Viermetz et al., 2008; Zhou et al., 2017)- and the determinants of why some topic lineages extend longer while others fall short.

Online social networks - such as Twitter, Facebook, or Reddit provide a valuable resource to study human behavior at large scale (Salganik, 2019), 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 (Garimella & Weber, 2017), information diffusion (Stai et al., 2018), the spread of disinformation (Murayama et al., 2021) and information virality (HOANG et al., 2011). While OSNs provide access to large datasets, they come at the expense of requiring pre-processing and feature engineering to be studied (Abidin et al., 2019; Gani & Chalaguine, 2022), of underlying biases that need to be accounted for (Yang et al., 2022), and experiments that need to be designed (Tan et al., 2014). Many techniques have become popularly adopted to address such challenges: text-mining and machine learning methods have been used to estimate the Sentiment (Li & Wu, 2010; Medhat et al., 2014; Redhu et al., 2018), Morality (Araque et al., 2020; Hopp et al., 2021; Johnson & Goldwasser, 2018), or Toxicity (Georgakopoulos et al., 2018) load in speech; network analysis (Grandjean, 2016) 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.

2. RELATED WORK

The advent of Online Social Networks (OSNs) has revolutionized the way we study human behavior and social dynamics. These platforms have become a treasure trove of observational data, providing researchers with unprecedented access to real-time, large-scale social interactions. Among various OSNs, Twitter, a microblogging online social network, has attracted significant attention from the academic community. The unique features of Twitter, including its 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, make it a valuable platform for academic research.

One of the key areas of research in this context is topic modeling. Topic modelling is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents, essentially extracting the hidden thematic structure in an unstructured set of texts. In the context of Twitter, it can be used to identify sets of tweets that share a common vocabulary, thereby providing insights into the prevalent discussions or themes on the platform. However, applying topic modeling to Twitter data presents several challenges. Due to the brevity of tweets, there is often a sparse co-occurrence of words across documents. Additionally, the informal language, high content variability, and the presence of noisy and irrelevant data can affect the accuracy and reliability of topic models. As highlighted by Federico Albanese and Esteban Feuerstein in their paper, "Improved Topic modeling in Twitter through Community Pooling," 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 (Albanese & Feuerstein, 2021). Their work proposes a novel pooling scheme for topic modeling in Twitter, which groups tweets whose authors belong to the same community on a user interaction graph, thereby addressing some of these challenges and enhancing the effectiveness of topic modeling on Twitter.

Building on the concept of topic modeling, a more nuanced approach considers the temporal dimension, leading to the development of dynamic topic modeling. Dynamic topic modeling allows for the tracking of the evolution of discourse over time, which is particularly relevant in the context of Twitter due to the timestamp associated with each tweet. This temporal aspect enables researchers to observe how topics evolve, merge, split, or disappear over time, providing a more comprehensive understanding of the discourse dynamics. Several researchers have explored dynamic topic modeling in their work. For instance, Jinjin Guo, Longbing Cao, and Zhiguo Gong, in their paper "Recurrent Coupled Topic Modeling over Sequential Documents," propose a model that assumes a current topic evolves from all prior topics with corresponding coupling weights, forming the multi-topic-thread evolution (Guo et al., 2021). This approach provides a more nuanced understanding of topic evolution, taking into account the influence of past topics on the current discourse. Similarly, the paper "Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. governors and cabinet executives" by Hao Sha, Mohammad Al Hasan, George Mohler, and P. Jeffrey Brantingham applies a dynamic topic model to COVID-19 related tweets by U.S. Governors and Presidential cabinet members (Sha et al., 2020). Their work tracks evolving sub-topics around risk, testing, and treatment, demonstrating the practical application of dynamic topic modeling in understanding real-world discourse.

In the context of topic evolution, Xicheng Yin, Hongwei Wang, Pei Yin, and Hengmin Zhu, in their paper "Agent-based opinion formation modeling in social network: a perspective of social psychology," studied topic evolution from a sociological and psychological perspective (Yin et al., 2018). They propose an agent-based online opinion formation model based on attitude change theory, group behavior theory, and evolutionary game theory, providing a unique perspective on the dynamics of topic evolution. The study of community dynamics in relation to topic modeling on Twitter offers another layer of complexity and depth to our understanding of online discourse. Communities on Twitter, often formed around shared interests, ideologies, or affiliations, can significantly influence the evolution of topics and the overall discourse on the platform. The paper "Modeling community structure and topics in dynamic text networks" by Teague Henry, David Banks, Christine Chai, and Derek Owens-Oas, provides a novel perspective on this aspect (Henry et al., 2018). They propose a Bayesian method that allows topic discovery to inform the latent network model and the network structure to facilitate topic identification. This approach underscores the interplay between community structures and topic evolution, highlighting the importance of considering social dynamics in topic modeling.

In a similar vein, the paper "Topic Lifecycle on Social Networks: Analyzing the Effects of Semantic Continuity and Social Communities" by Kuntal Dey, Saroj Kaushik, Kritika Garg, and Ritvik Shrivastava, presents an analysis of topic lifecycles with respect to communities (Dey et al., 2018). They characterize the participation of social communities in the topic clusters and analyze the lifecycle of topic clusters with respect to such participation. Their work provides valuable insights into how community dynamics can influence the emergence, evolution, and disappearance of topics on Twitter. Another significant contribution to this field is the paper "Online Tensor Methods for Learning Latent Variable Models" by Furong Huang, U. N. Niranjan, Mohammad Umar Hakeem, and Animashree Anandkumar (Huang et al., 2015). They introduce an online tensor decomposition-based approach for community detection and topic modeling, considering decomposition of moment tensors using stochastic gradient descent. This approach offers a novel way of understanding community dynamics and their influence on topic modeling. The exploration of toxicity in online discourse and its dynamics is a critical aspect of research in online social networks, particularly on platforms like Twitter. The presence of toxic speech, characterized by hateful or harmful language, can significantly influence the nature of discourse and the evolution of topics on these platforms. A noteworthy contribution to this field is the paper "Hate is the New Infodemic: A Topic-aware Modeling of Hate Speech Diffusion on Twitter" by Sarah Masud, Subhabrata Dutta, Sakshi Makkar, Chhavi Jain, Vikram Goyal, Amitava Das, and Tanmoy Chakraborty (Masud et al., 2020). Their work focuses on exploring user behavior that triggers the genesis of hate speech on Twitter and how it diffuses via retweets. Their findings provide valuable insights into the mechanisms of hate speech propagation and the role of topic popularity in this process.

The study of toxicity dynamics in topic evolution is a relatively new area of research, and these works provide a foundation for further exploration. Understanding these dynamics can inform interventions and policy-making in addressing toxic behavior in digital communities, making it a crucial aspect of research in online social networks.

The study of toxicity in online discourse is not only about identifying its presence but also about understanding its nuances and variations. This requires sophisticated text mining approaches that can detect properties like toxicity or the moral load of a document. These approaches are typically based on a corpus of human-labeled documents, where each document is transformed into a feature vector that embeds the relevant properties of the document. Machine learning is then applied to train a model on the feature vectors using the human-generated annotations as the ground truth.

One of the significant challenges in this area is the complexity and variability of human language. Toxic speech can take many forms, from explicit hate speech to more subtle forms of harmful language. Therefore, the text mining approaches need to be robust and sophisticated enough to capture this complexity and variability. The paper "A Bayesian Nonparametric Latent Space Approach to Modeling Evolving Communities in Dynamic Networks" by Joshua Daniel Loyal and Yuguo Chen provides a valuable perspective on this aspect (Loyal & Chen, 2020). They present a Bayesian nonparametric model for dynamic networks that can model networks with evolving community structures. Their model extends existing latent space approaches by explicitly modeling the additions, deletions, splits, and mergers of groups with a hierarchical Dirichlet process hidden Markov model. This approach provides a robust framework for detecting properties like toxicity in online discourse.

The unique contribution of this research lies in the development of a dynamic topic evolution model, a novel approach designed to analyze the intricate relationship between the evolution of topics and the manifestation of speech toxicity within geolocated tweets. This innovative model is a significant departure from traditional static topic modeling approaches, which typically fail to capture the temporal dynamics inherent in social media discourse.

Our model leverages the power of sBERT (sentence-BERT) sentence embeddings, a modification of the widely-used BERT model specifically designed for sentence-level applications (Reimers & Gurevych, 2019). sBERT embeddings capture the semantic meaning of tweets, enabling a more nuanced understanding of the topics being discussed. By representing the topic evolution model as a directed graph, we can effectively track and visualize the progression of topics over time, and crucially, the corresponding changes in speech toxicity. To ensure the robustness and validity of our model, we introduce a hashtag-based method for validation and hyperparameter selection. Hashtags, often used in tweets to denote the central theme or sentiment, provide a valuable source of ground truth for validating our topic evolution model. By analyzing the co-occurrence of hashtags and topics, we can fine-tune the model's hyperparameters to optimize its performance.

This research significantly expands on the existing literature by addressing a critical gap in our understanding of the dynamics of speech toxicity in topic evolution. Prior studies have largely treated topic evolution and speech toxicity as separate phenomena. Our work, however, underscores the interconnected nature of these two aspects of online discourse. By examining them in tandem, we can gain a more comprehensive understanding of how topics evolve and how this evolution can influence, or be influenced by, the level of toxicity in the discourse. The implications of this research extend beyond academia. By shedding light on the dynamics of speech toxicity in online discourse, we can inform the development of more effective moderation tools for digital communities, contribute to the creation of healthier online spaces, and provide valuable insights for policymakers seeking to regulate online hate speech and harassment.

3. DATA AND METHODS

In this research, we employ a comprehensive dataset composed of approximately eight million tweets harvested from June 2, 2020, to November 3, 2020. These tweets were geographically pinpointed around the metropolitan area of New York City, USA. The selected timeframe for data collection coincides with a significant and tumultuous period in recent American history, characterized by several high-profile events.

Specifically, this period witnessed the nationwide demonstrations ignited by the tragic death of George Floyd, a Black man, in police custody. This incident sparked unprecedented civil unrest and breathed new life into the Black Lives Matter (BLM) movement, an activist group that campaigns against systemic racism and violence towards Black individuals. Simultaneously, the 2020 presidential election race was unfolding, an event that polarized the country and amplified social discourse on a national scale. The collected dataset was sourced utilizing the academic Twitter API, an open-access resource provided by Twitter for scholarly research. This API facilitates large-scale data collection, offering a substantial amount of publicly available information to researchers. In this case, the only imposed restriction was the geographical location parameter, which ensured the data was exclusive to New York City's vicinities. Distinct from other related studies that primarily focus on specific topics or individual accounts, our research adopted a broader approach. We resorted to the collection of a random assortment of tweets from diverse users and subjects, assembled in bulk from the live Twitter timeline. This methodology enables an expansive view of the public discourse during the chosen timeframe. The non-specific selection of tweets offers a more authentic reflection of the general sentiment and diverse conversations happening on the platform, rather than focusing on a specific conversation or set of accounts. This is especially critical considering the momentous societal and political events occurring during the specified period, providing us with valuable insight into real-time public responses and perceptions.

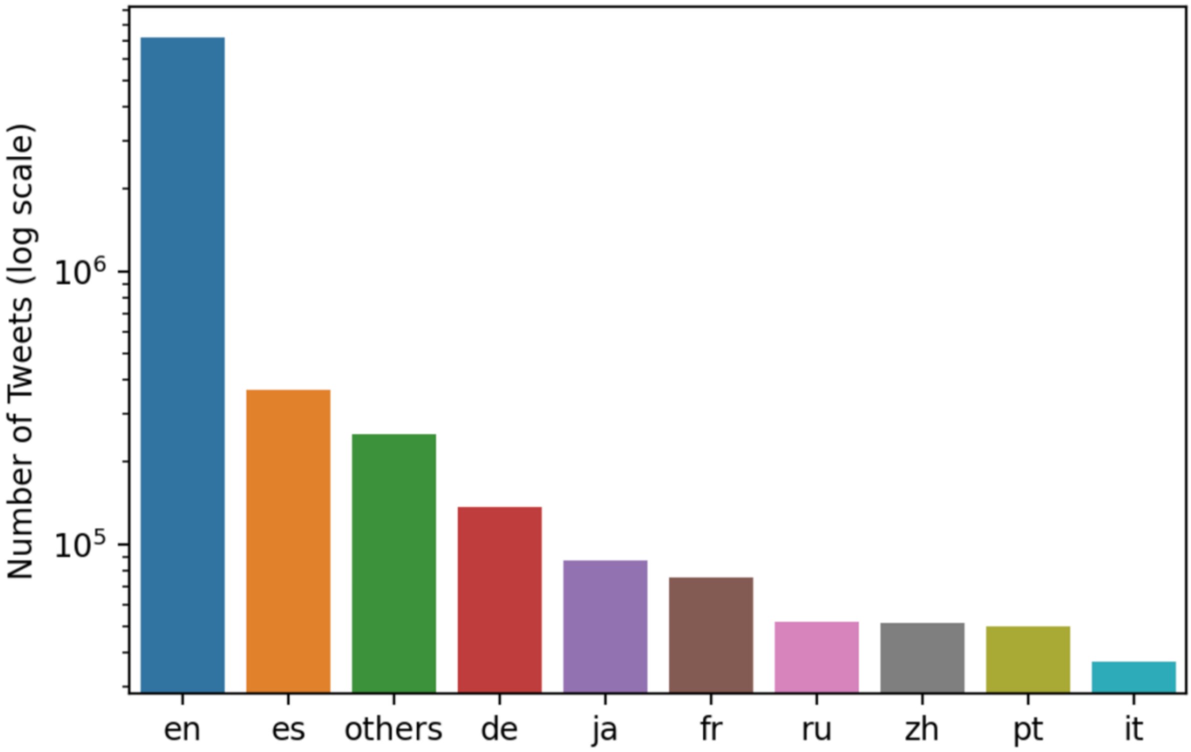

Figure 1 - Language of collected Tweets as classified by the fasttext language detector

Our dataset of tweets exhibits a uniform distribution over time, averaging 53,859 tweets per day over a span of 154 days. The corpus comprises a total of 364,918 unique hashtags, with each hashtag surfacing an average of 4.81 times throughout the dataset. This metric provides an indication of the thematic diversity present within the data, highlighting the multitude of conversations and topics encompassed in the dataset.

To counteract data sparsity and to bolster the frequency of token co-occurrences - a key metric for establishing correlations and similarities between documents - we instituted a series of preprocessing measures across the entire tweet corpus. These steps are designed to standardize the data and eliminate extraneous elements that do not contribute to semantic understanding, thereby refining the quality of the dataset for subsequent analysis.

Firstly, we converted all alphabetical characters to lowercase. This helps to maintain consistency, reducing the chance of identical words being treated as separate entities due to differences in case. Secondly, we removed all emojis from the dataset. While emojis can sometimes carry sentiment and contextual clues, their interpretation can vary widely among users and cultures, potentially introducing noise into our analysis. Deduplication of tweets was another crucial step to ensure that each data point is unique, removing redundant information that could skew the analysis and interpretation of results. We also removed all non-alphanumeric characters, stripping out punctuation and special characters that do not typically contribute to the semantic content of the text. Further, we discarded tokens containing fewer than two characters and filtered out tweets comprising less than 20 characters. These measures aim to ensure that our analysis focuses on meaningful textual content, reducing the influence of brief, potentially out-of-context data points. The fastText language detector (Joulin et al., 2016) was then employed to filter out tweets not written in English. Given that our analysis is anchored in English-language context, this ensures that all analyzed tweets fall within the scope of our linguistic understanding. Lastly, we implemented lemmatization, a natural language processing technique that transforms words into their base or dictionary form (e.g., "running" becomes "run"). This allows us to consolidate different grammatical forms of a word, enhancing our ability to identify patterns and associations within the data.

Following the application of our described preprocessing steps, the size of our tweet corpus was significantly reduced. A reduction of 37% culminated in a refined dataset of 5,197,172 tweets. This diminution is due in part to the exclusion of non-English tweets, removal of duplicates, and the discarding of brief and potentially less meaningful content.

The core focus of this research is to analyze the evolution of topics within our defined time period. To facilitate this, we partitioned the tweet corpus into distinct, non-overlapping time intervals, each spanning ten days. This partitioning resulted in roughly 350,000 tweets per interval, yielding a total of eleven discrete time intervals. The choice to use non-overlapping time intervals, as opposed to a sliding window method, was a conscious decision rooted in practical and analytical considerations. A sliding window method, which includes overlapping time frames, could potentially inflate the number of time intervals and subsequently increase computational demands, posing a challenge for resource allocation and data processing efficiency. Moreover, non-overlapping time intervals offer a clearer chronological progression and distinct temporal segments, which aids in tracking the emergence, persistence, and dissipation of topics over time. This enables more straightforward tracing of topic evolution without the possibility of topic dilution that could occur due to overlap in the sliding window method. The clarity gained from this approach allows for more efficient analysis and stronger conclusions about the development of conversation themes on Twitter within the chosen period.

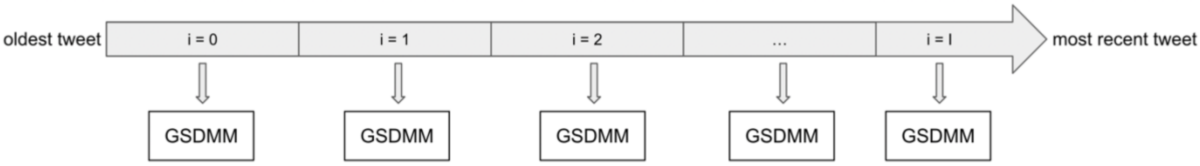

Figure 1 - Time Binning of Tweet Corpus

3.1. Sentence-BERT Embedding Model

In the realm of Natural Language Processing (NLP), Sentence-BERT (SBERT) stands as a key advancement tailored to meet the needs of semantic similarity tasks. SBERT is an enhancement of the well-established Bidirectional Encoder Representations from Transformers (BERT) model, specifically engineered to overcome the computational inefficiencies of BERT when dealing with large datasets. Utilizing a novel approach of siamese and triplet network architectures, SBERT crafts semantically-rich sentence embeddings that are more efficient to compare using cosine similarity - an essential aspect for semantic similarity applications.

This innovation is achieved via the incorporation of pooling operations on the output of BERT, resulting in fixed-size sentence embeddings suitable for comparison. Several pooling strategies are available, such as using the output of the 'CLS' token, or calculating the mean or max-over-time of all output vectors. The BERT model is fine-tuned through these siamese and triplet network structures, resulting in semantically potent sentence embeddings that can be compared efficiently using cosine-similarity. The training of SBERT is executed on a combination of the Stanford Natural Language Inference (SNLI) and Multi-Genre Natural Language Inference (MultiNLI) datasets. These datasets contain over a million sentence pairs annotated with labels like 'contradiction,' 'entailment,' and 'neutral,' providing a rich training base. Once trained, the efficacy of SBERT is assessed on semantic similarity tasks, where it outperforms the original BERT and RoBERTa models, providing a robust and efficient solution for semantic similarity applications.

3.2. Topic Modelling and DAG Lineage

In our study, we employed the Generalized Scalable Dirichlet Multinomial Mixture Model (GSDMM) for topic identification within our tweet corpus. As an extension of the widely used Latent Dirichlet Allocation (LDA) model, GSDMM is a preferred choice for tackling sparse data, making it particularly suitable for our refined dataset. The GSDMM was run independently for each of the non-overlapping time bins, as depicted in Figure 2. Subsequent to this, we executed a process of topic linkage across the different time intervals. This involved establishing connections between common or related topics identified in different time windows, providing an approximation of topic evolution over the studied period and delineating their lineage.

Unlike LDA, which assumes that documents contain a mixture of topics, GSDMM operates under the assumption that a document, or in this case a tweet, typically covers one single topic. This makes it an excellent choice for shorter, more focused text documents, such as tweets. The GSDMM model necessitates the specification of two hyperparameters: α and β. The α parameter represents the relative weight given to each cluster of words, while β designates the significance of each word in determining a document's topic distribution. Following the guidance of Yin and Wang, we chose values of 0.1 for both α and β.

To ensure the robustness of our findings, we ran GSDMM for 20 iterations for each time bin. One of the salient advantages of GSDMM is its ability to infer the number of topics automatically. Hence, for our analysis, we merely needed to initialize the number of topics, which we set at 120. By allowing GSDMM to deduce the number of topics, we aimed to capture the organic complexity and nuance present in public discourse on Twitter during the studied period. At the conclusion of the aforementioned step, each of our temporal segments, or time bins, is treated as a separate entity, and the topics identified within each are currently unlinked. Each tweet in our dataset can be traced to a specific time bin and associated with a particular topic.

To maintain uniformity in our notations throughout the manuscript, let's establish the following convention: M_j^i will denote a set of tweets that are linked with topic 'j' at time 'i'. This ensures a clear and consistent representation of our data, making the analytical process more straightforward and intuitive.

Moving forward to the aspect of topic evolution, our approach leverages the concept of a Directed Acyclic Graph (DAG). The Graph is partitioned (M-partite), with nodes representing the topics and directed edges signifying their progression over time. Each partition within this graph corresponds to all topics j from a specific time bin i.

The edges in this graph, which connect the topics, are interpreted as an ordered pair of nodes (M_j^i, M_j^{i+1}) in two consecutive time windows. This ordering signifies a 'parent-child' relationship between topics in successive time intervals, illustrating the lineage and evolution of topics over time. For example, an edge from node M_j^i to M_j^{i+1} indicates that topic j at time i has evolved or carried forward into the same topic j at time i+1.

Such a representation allows us to visualize and analyze the dynamics of topics over time, revealing how conversations on Twitter evolve and adapt in response to unfolding events and changing public sentiments. It helps us in tracing the journey of specific topics, thereby providing a comprehensive understanding of the discourse during the studied period.

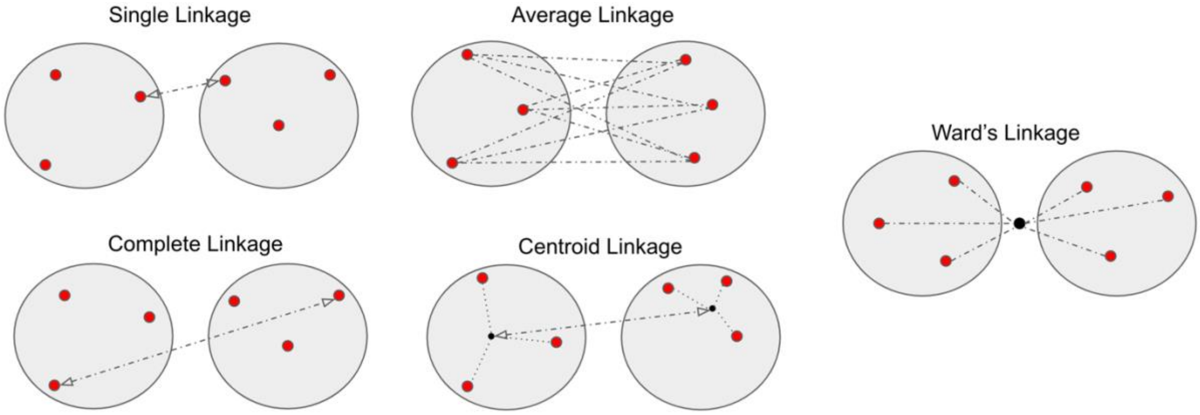

Figure 2 - Collection of Linkage Methods that can be used to quantify the distance between two Topics. Please note that it is required to have an embedding on the document level. In our case, we utilize sBERT to obtain a numerical vector for each Tweet.

In our quest to identify the relationships and continuity between topics across time intervals, we assess the semantic similarity between all topics in successive partitions of our Directed Acyclic Graph (DAG). To accomplish this, we employ SentenceBERT (sBERT), a variant of the popular transformer-based BERT model that has been pre-trained specifically for sentence-level tasks.

Each tweet in our corpus is encoded by sBERT into a high-dimensional feature vector. This process transforms the textual content of tweets into numerical representations that encapsulate their semantic content. The cosine distance between any pair of such vectors serves as a quantifier of inverse semantic similarity, i.e., a smaller distance corresponds to higher semantic similarity between the tweets, and vice versa.

When a topic is defined as a set of data points (tweets) in this high-dimensional vector space, the proximity or similarity between two topics is gauged by comparing their respective sets of data points. For this purpose, we utilize the Centroid Linkage approach, inspired from the field of hierarchical clustering. In the Centroid Linkage approach, the distance between two clusters (or topics, in our case) is defined as the distance between their respective centroids (the mean vector of the data points within a cluster). We preferred this method over alternatives like single or complete linkage methods, owing to its robustness against outliers - it is less influenced by extremely close or far data points.

Furthermore, the Centroid Linkage approach is also more computationally efficient than the Ward linkage method, which involves more complex calculations to minimize the total within-cluster variance. Given the scale of our dataset and the high-dimensional nature of our vectors, computational efficiency was a significant consideration in our choice of methods. The combination of sBERT encoding and Centroid Linkage thus provides us a robust and efficient methodology to track topic progression, allowing us to identify the threads of conversation as they evolve and shift over time.

To further refine the insights derived from our dataset, we implement a threshold-based method to establish which potential connections, or candidate edges, between topics in our Directed Acyclic Graph (DAG) should be included.

This method relies on the notion of semantic similarity, measured by the cosine similarity, which quantifies the degree of relatedness between two topics based on their semantic content. A higher cosine similarity value indicates a higher degree of relatedness. We apply a threshold parameter, represented as ε, to this measure. In effect, we only include edges in our DAG that exhibit a cosine similarity value greater than or equal to ε. This process allows us to focus on substantial topic continuities, emphasizing the main threads of discourse over the studied period.

In this manner, we build a precise representation of topic progression, tracing the most significant semantic threads and highlighting key shifts and continuities in the public discourse on Twitter.

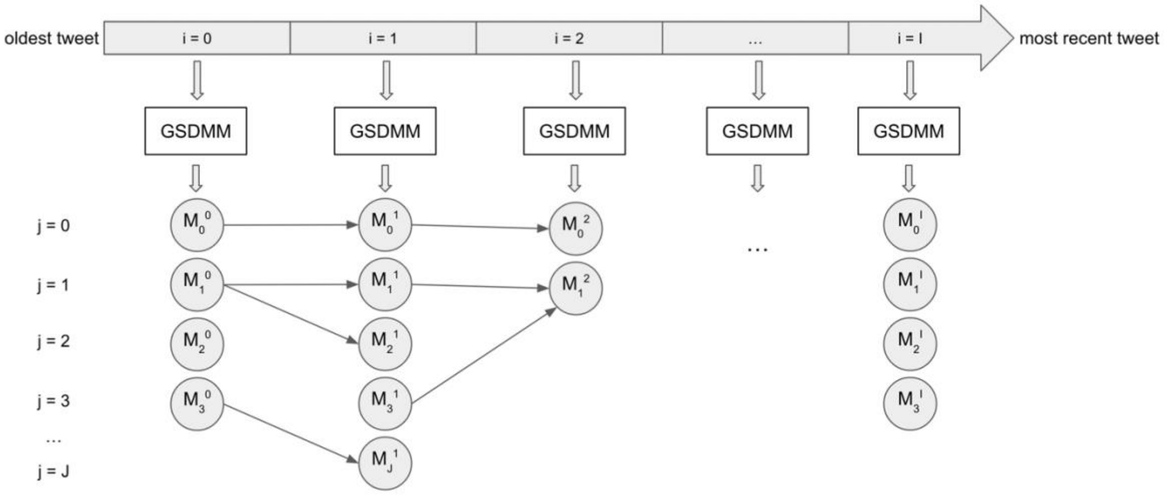

Figure 3 - Static Topic results are used to build a graph representation of the Topic Evolution. Topics of adjacent Time Bins are connected if a proximity threshold ε is exceeded.

3.3. GSDMM Text Clustering

The Generalized Scalable Dirichlet Multinomial Mixture Model (GSDMM) is a text clustering algorithm that is uniquely suited to handle short text documents (Udupa et al., 2022). In this detailed overview, we delve deeper into its process, which includes three main steps: initialization, iteration, and convergence.

Initialization: GSDMM commences by defining a predetermined number of topics, which, for all intents and purposes, function as clusters. It then arbitrarily assigns each document from the dataset to one of these topics. For instance, if we set the number of topics as ten for a dataset comprising tweets, every tweet will be randomly ascribed to one of these ten topics. This stage essentially forms the baseline for the iterative process that follows.

Iteration: In the iterative phase, GSDMM systematically reviews each document, assessing the need to reassign it to a different topic. The likelihood of a document being relocated to another topic is reliant on two crucial factors.

- The prevalence of documents within a topic: This factor essentially encourages the growth of larger, more comprehensive topics. If a topic already houses a significant number of documents, a new document is more likely to be categorized under that topic. It is based on the rationale that documents sharing similarities would collectively form more robust and broader topics.
- The correlation between the word distribution within the document and the topic: This factor promotes the coherence of topics. If the document's word distribution closely mirrors the topic's word distribution, it indicates a higher degree of semantic relevance, hence making it more likely for the document to be assigned to that topic.

This iterative process is repeated, cycling through every document in the dataset.

Convergence: The iteration phase continues until we reach the stage of convergence, which is the point where the allocation of documents to topics becomes relatively stable. Stability, in this context, implies that further iterations would not lead to substantial changes in document assignments. Upon reaching this stage, the algorithm halts. As a result, each topic can be interpreted as a cluster of documents with close semantic ties.

It is worth noting that this algorithm particularly excels in dealing with 'short texts' (like tweets), which typically do not possess a clear topical structure that longer documents might have. Its utility lies in its ability to intuit semantic connections and group documents into distinct, coherent clusters, even in the absence of strong, overt topic indicators.

3.4. Detection of Transition Types

As we explore the semantic journey of topics through time, our methodology allows for dynamic interconnections between topics across adjacent time bins. By definition, a topic, represented as M_j^i , can be linked to multiple successor topics in the next time bin (i + 1). This occurs if several topics in time bin (i + 1) bear a similarity with M_j^i that surpasses the threshold ε. Consequently, a parent topic node in our Directed Acyclic Graph (DAG) can branch out into multiple child topics.

The converse is also true: a child topic node can have multiple parent topics. If several topics from time bin i are similar enough to a topic in time bin (i + 1) such that they exceed the ε threshold, this forms a many-to-one connection from one time bin to the next.

This configuration results in five possible transition scenarios between two adjacent time bins, which we refer to as 'transition types' in our study:

- Topic Splitting: If a topic M_j^i from time bin i is associated with more than one topic in the succeeding time bin (i + 1), this means the original topic has "split" into multiple child topics.

This condition is observed when the outdegree (number of outgoing edges from a node) of M_j^i exceeds 1.

- Topic Merging: If more than one topic from time bin i shares a substantial semantic similarity with a single topic M_j^{i+1} in time bin (i + 1), this indicates multiple parent topics have "merged" into one child topic. This is observed when the indegree (number of incoming edges to a node) of M_j^{i+1} exceeds 1.
- Topic Stagnation: A topic M_j^i is considered to "stagnate" if it has only one child topic M_j^{i+1}, and this child topic in turn has only one parent topic. This scenario occurs when M_j^i has an outdegree of 1, and its child topic has an indegree of 1.
- Topic Disappearance: If a topic M_j^i does not have any semantically similar topics in the following time bin, i.e., it has an outdegree of 0, we denote this topic as having "disappeared".
- Topic Emergence: Conversely, if a new topic M_j^{i+1} in time bin (i + 1) does not have any semantically similar topics in the preceding time bin, indicated by an indegree of 0, we describe this scenario as a topic "emergence".

These five scenarios are referred to as transition types throughout the rest of this paper.

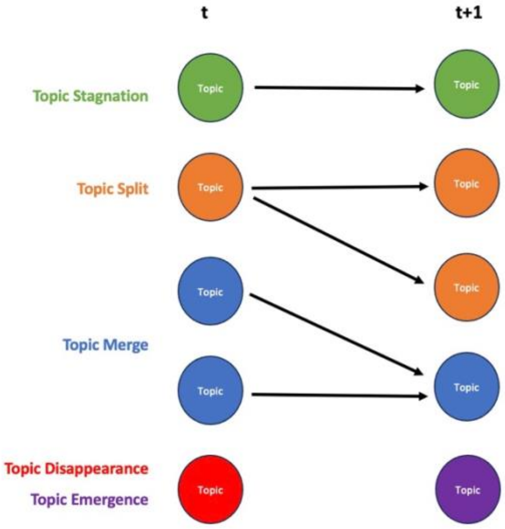

Figure 4 - Definition Transition Types

We define that topics situated in the first time bin of our analysis can't be classified as "emerging" topics. This is due to the fact that we don't have any preceding data from which these topics could potentially emerge. The semantic content in the first time bin serves as our initial observation point, so all topics in this bin are treated as baseline topics, not emergent ones.

Similarly, topics found in the last time bin of our dataset cannot "disappear" by our definition. Given the absence of any subsequent data beyond this point, we don't have the means to observe or quantify their disappearance. Instead, these topics form the terminal points of our analysis, representing the state of discourse at the end of our observation period.

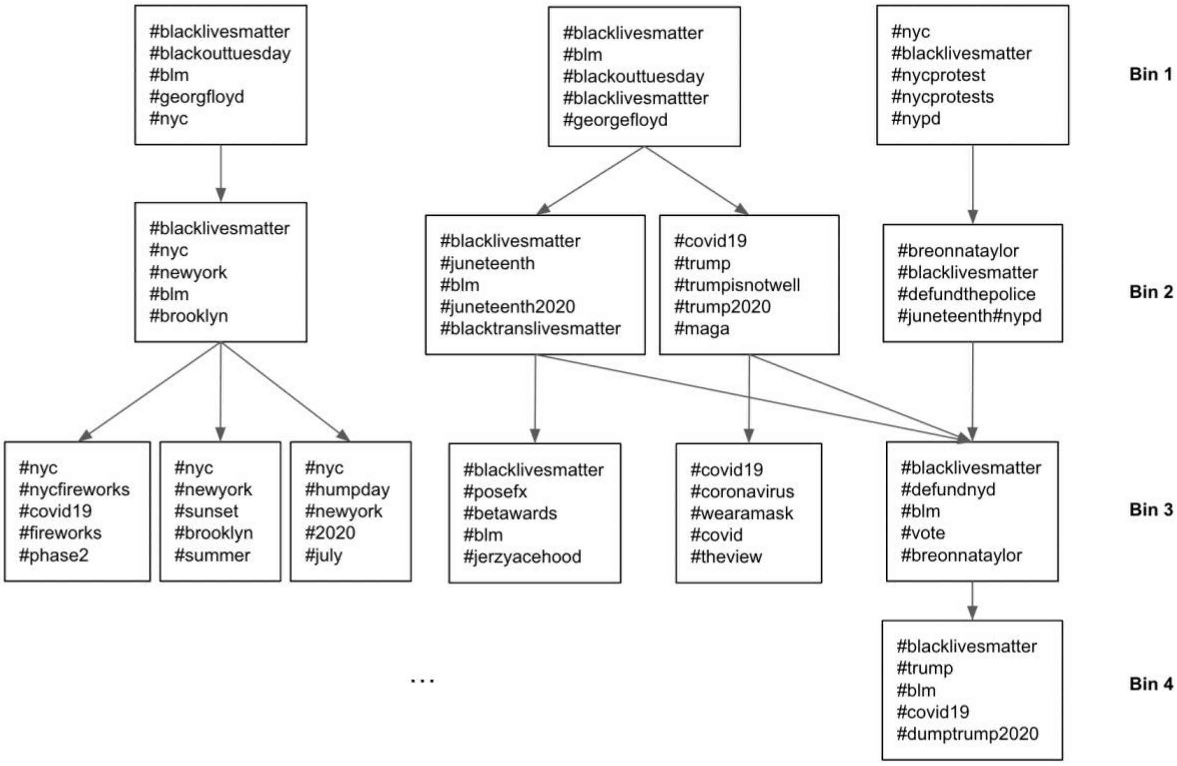

Figure 5 - Sample Topic Trajectories. Topics labeled with the 5 most frequent hashtags. A proximity threshold ε = 0.8 was chosen.

Figure 5 serves as an illustrative example of our method of tracing the evolution of a specific topic across time bins. In this instance, we track the lineage of topics linked to the Black Lives Matter (BLM) movement across four time bins.

In this visualization, each box represents a distinct topic, with the five most frequently used hashtags listed therein. Topics within the first time bin associated with the BLM movement primarily connect to subsequent topics which also feature BLM-related hashtags among their most frequently used.

An intriguing case of topic evolution can be observed between Time Bin 1 and Time Bin 2. The BLM topic appears to undergo a split, diverging into two distinct threads. One thread continues to focus primarily on the BLM protests and related events. The other, however, takes a more political turn, incorporating discussions around then-president Donald Trump. This split offers a tangible example of how social conversations on Twitter can diversify, often weaving in different but related strands of public interest and concern.

This analysis, using user-generated hashtags as a form of 'ground truth', validates the consistency of our GSDMM-based topic detection model. The alignment between the hashtags and the identified topics shows our model's accurate tracing of topic evolution, providing a valuable verification of our methodology.

The hyperparameter ε, serving as a threshold for semantic similarity, profoundly impacts the resulting structure of the Topic Evolution's Directed Acyclic Graph (DAG). The graph's configuration and thereby the patterns of topic transitions are sensitive to this parameter.

To understand this more clearly, consider the extreme ends of ε 's spectrum. If ε is chosen to be close to its maximum, indicating a stringent similarity requirement, only highly identical topics across adjacent time bins will be linked. In this scenario, most topic transitions would be classified as 'emergence' and 'disappearance', due to the strict proximity requirement resulting in very few detections of child or parent topics.

On the other hand, if ε is chosen near its minimum value, a lenient similarity threshold is set, connecting almost all topics to those from adjacent time bins. Consequently, there would be a surge in 'splitting' and 'merging' transition types, as topics loosely connect with a multitude of other topics between time bins.

Figure 6 elucidates the sensitivity of ε by exhibiting the proportion of each transition type at various ε values. This clearly illustrates the significant influence of the hyperparameter ε on the complexity and topology of the evolving topic network. Such understanding aids in the judicious selection of ε, balancing the trade-off between network complexity and semantic continuity.

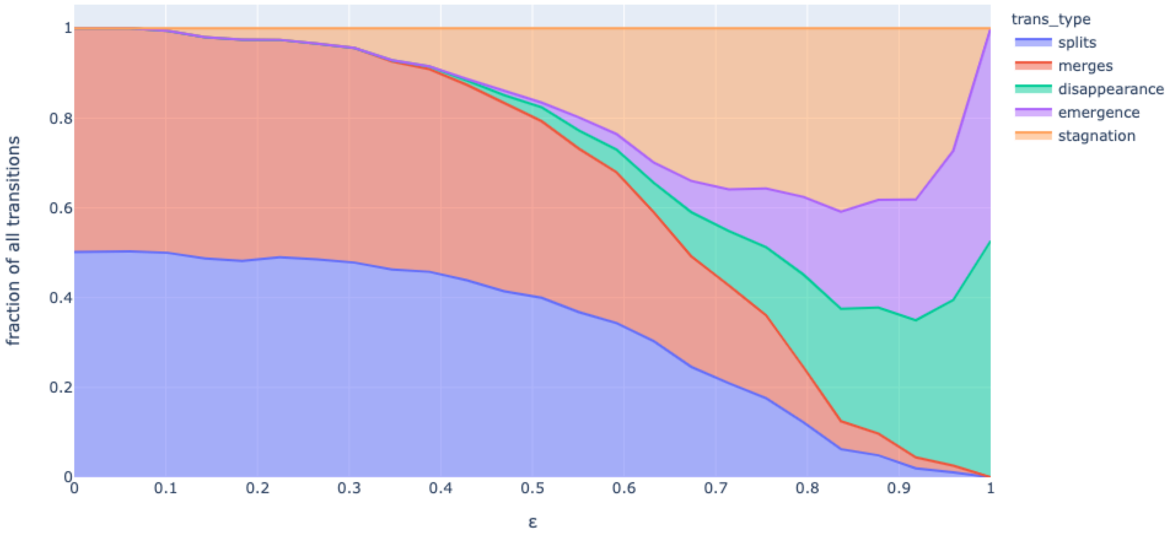

Figure 6 - Impact of the proximity threshold ε on the Transition Types Distribution

The validation of our dynamic topic evolution Directed Acyclic Graph (DAG) is accomplished by examining the congruity between the linkage inferred from BERT text embeddings and the distribution of hashtags across topics. Hashtags, being user-generated, can arguably serve as practical, ground truth labels for tweet topics. In this context, we utilize the Jensen-Shannon Divergence to quantify the similarity of hashtags for each edge in the DAG, providing a complementary, data-driven measure of topic continuity.

When we commence with a threshold ε near 1, our method connects only very similar topics. This leads to a sparsely linked DAG, where the few connected topics exhibit similar hashtag distributions. As we decrease ε, the topics linked become less similar, which, in turn, increases the number of edges in the DAG. Correspondingly, the similarity of hashtag distribution among connected topics diminishes, as broader, less similar topics are linked.

This pattern, where hashtag similarity of connected topics increases with ε, serves as empirical confirmation of the model's consistency. The model's ability to capture this intuitive relationship provides further assurance that our approach to topic linkage, grounded in the high-dimensional semantic space created by BERT, resonates well with the practical, user-defined context of hashtags. This underlines the efficacy of our approach in accurately tracking the evolution of topics over time.

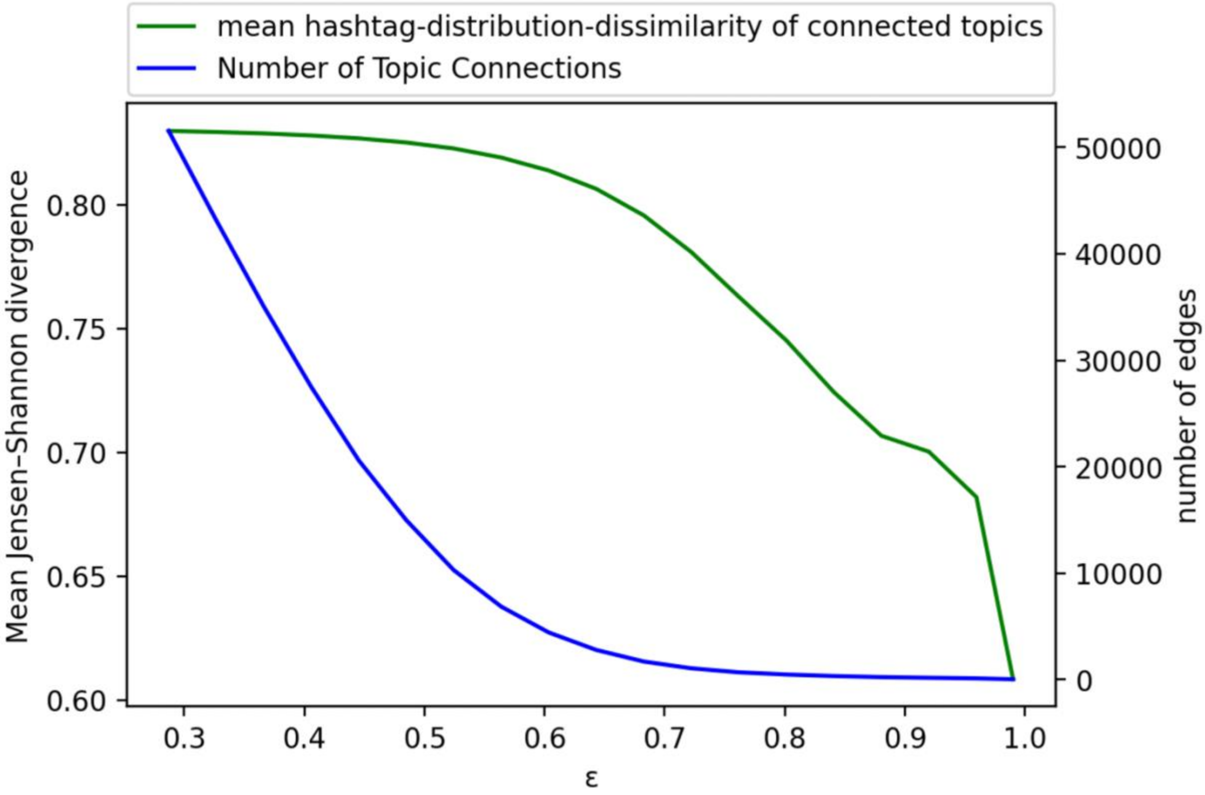

Figure 7 - Similarity of the distribution of hashtags across connected Topics in different values for ε

4. RESULTS: EXPLORING TOXICITY IN TOPIC EVOLUTION

In our examination of the language present within our text corpus, we pay particular attention to 'toxicity'—the presence of harmful or inflammatory language that could potentially trigger negative reactions or instigate conflict. Laura Hanu (Hanu & Unitary team, 2020) developed an advanced machine learning model named 'Detoxify' that is capable of quantifying the level of toxicity embedded within a text.

Detoxify has been trained on a dataset consisting of human-annotated Wikipedia comments. While these comments and Twitter tweets differ in various aspects, including length, context, and user demographics, the model's architecture and learning process are presumed to afford it a degree of generalizability. Therefore, despite not being directly trained on tweets, Detoxify is expected to offer reliable estimates of toxicity levels within our Twitter text corpus.

It's important to note that Detoxify is a multi-label model, meaning it can provide estimates for multiple attributes of speech, ranging from obscenity and insult to threat and identity-based hate. However, for the scope of our analysis, we are specifically interested in the toxicity measure, which encapsulates the general level of harmful or inflammatory language. By integrating Detoxify's toxicity estimates with our topic evolution analysis, we aim to explore how toxicity levels change and propagate across different topics and time periods in our dataset.

The toxicity measure returned by Detoxify is continuous, ranging from 0 (non-toxic) to 1 (highly toxic). However, a visual inspection of this distribution reveals a bimodal pattern, prompting us to transform these continuous values into a binary variable. This dichotomization aligns with the model's original training data, which relied on discrete, human-annotated labels, making our analysis more consistent and the interpretation of results more straightforward.

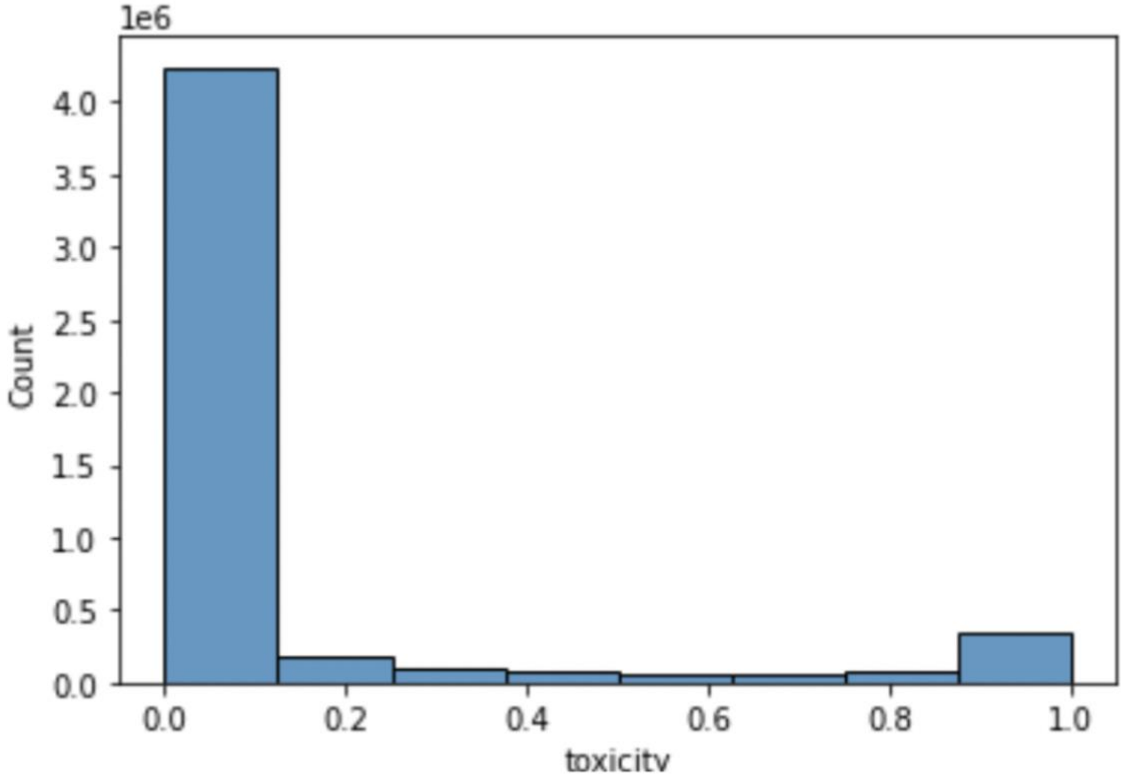

Figure 8 - Histogram showing the bimodal distribution of the embedded Toxicity in tweets.

By implementing a cut-off threshold at 0.5, we identify approximately 11% of tweets within our corpus as toxic. This proportion aligns with findings from other studies conducted within the domain of social media discourse analysis. It's important to note that the percentage of toxic tweets can vary based on the focus and the definition of toxicity used in the study. 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 a lower percentage. For instance, a study (Davidson et al., 2017) analyzed a dataset of over 25,000 tweets and found that around 5% of them contained hate speech. Another study (Founta et al., 2018) 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.

In the initial phase of our analysis, we delve into a general examination of tweet toxicity, considering its variance across topics and geographic space. We leverage the geolocation metadata attached to each tweet, focusing specifically on the district in New York City where the tweet originated. Prior to integrating the toxicity data into the temporal topic evolution graph, we believe it's helpful to present an overview of how tweet toxicity is distributed across both topics and space.

Two primary topics dominate the corpus - Black Lives Matter (BLM) and Covid-19. For each New York district, we compared the average toxicity of tweets pertaining to these two topics with the overall average tweet toxicity in that district. This approach controls for district-level variations in the baseline toxicity levels, which could potentially confound our results.

Essentially, a district might have a higher or lower average toxicity level due to unique local factors, such as differences in the political climate, demographic factors, or the impact of specific events. Therefore, to prevent inaccurate comparisons of toxicity levels between districts, we choose to make within-district comparisons, contrasting topic-specific toxicity with the district's average toxicity level.

Our findings indicate a striking difference in toxicity between the two chosen topics. Tweets related to BLM showed above-average toxicity in 10 out of the 11 districts studied, implying that discussions around this topic tend to have a higher degree of inflammatory language. On the other hand, tweets about Covid-19 displayed a below-average toxicity level in all 11 districts.

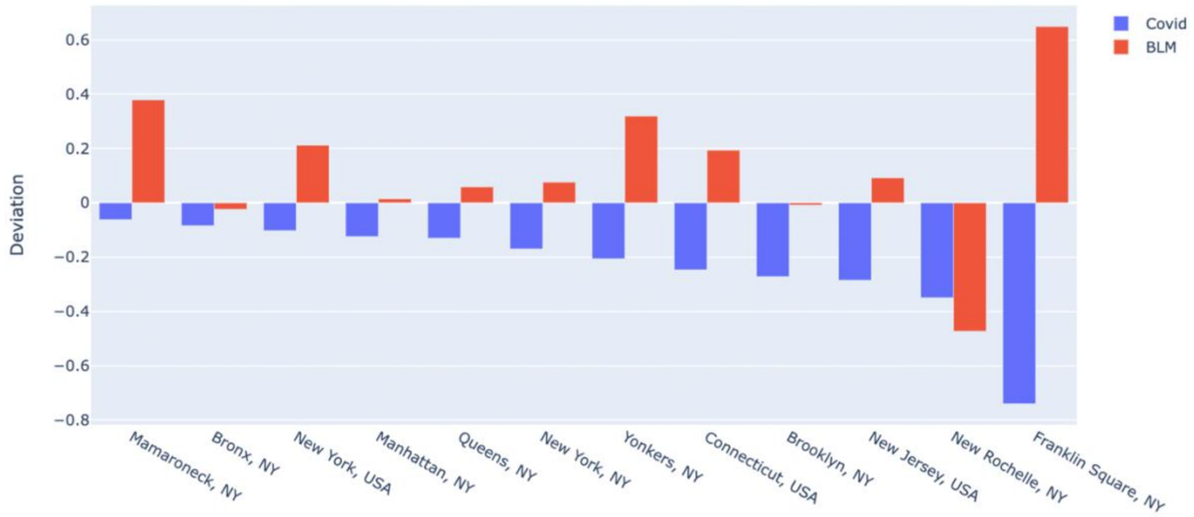

Figure 9 - Exploration of Tweet's Toxicity across Topics and Space - How much are Toxicity labels of BLM and Covid related Tweets deviating from the district's average Tweet Toxicity?

In the context of the Topic Evolution Directed Acyclic Graph (DAG), we can incorporate the binarized toxicity data to give us deeper insights into how toxicity changes as topics evolve. This integration process involves assigning an attribute to each topic node that represents the proportion of tweets categorized as toxic within that particular topic.

Furthermore, we introduce an attribute to each edge of the graph, termed $\Delta toxicity$, which reflects the relative change in the percentage of toxic tweets as we move from one topic to its successor in time. It quantifies how the toxicity evolves between a parent topic and its child topic.

The $\Delta toxicity$ is computed using the formula:

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

Here, Toxicity(childTopic) represents the proportion of toxic tweets in the child topic, while Toxicity(parentTopic) corresponds to the proportion of toxic tweets in the parent topic. A positive $\Delta toxicity$ value indicates an increase in toxicity from parent to child topic, while a negative value signifies a decrease.

4.1. Toxicity per Transition Type

In this section, we aim to explore the interplay between topic evolution, characterized by the five transition types identified earlier, and the toxicity level in tweets. Specifically, we investigate if the occurrence of a specific transition type tends to correlate with a significant change in the toxicity of the discourse, on average.

For each edge in the DAG, which corresponds to a transition between topics, we compute the $\Delta toxicity$ metric, reflecting the relative change in toxicity. As some transitions (such as merges and splits) involve multiple edges, they would, by definition, be associated with multiple $\Delta toxicity$ values.

Our interest lies in whether the transition types—namely, merges, splits, emergences, disappearances, or stagnations—exhibit distinct trends in average $\Delta toxicity$. This understanding could provide valuable insights into whether the process of topic evolution tends to incite or mitigate the prevalence of harmful language.

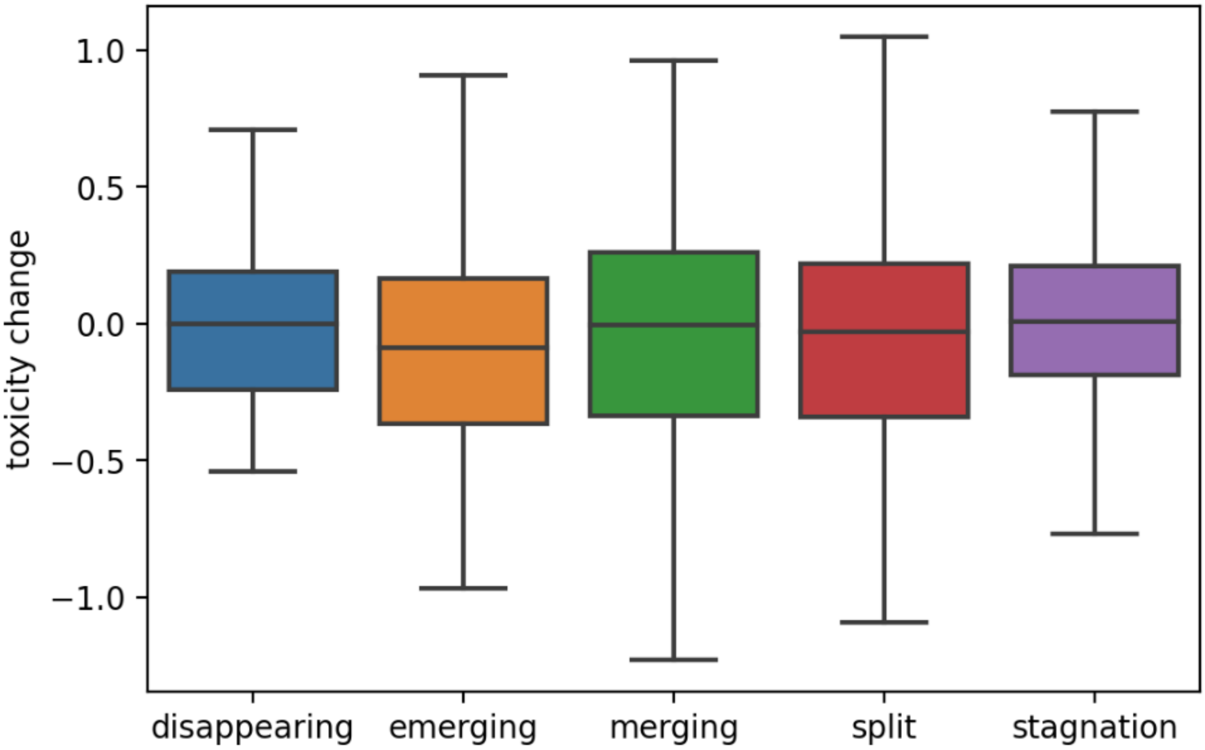

Figure 10 - Distribution of $\Delta toxicity$ broken down by transition type. A proximity threshold of ε = 0.8 was chosen. Topics with a Popularity < 30 were discarded.

The $\Delta toxicity$ distributions per transition type are visually represented in Figure 10. However, our analysis revealed no significant differences between these distributions. The mean $\Delta toxicity$ is approximately zero for all transition types, which suggests that no particular transition type is associated with a substantial shift in the percentage of toxic tweets.

In essence, the process of topic evolution in itself, characterized by the type of transition, does not appear to influence the overall toxicity level in the examined discourse. This absence of correlation indicates that toxicity in Twitter conversations is likely influenced by other factors beyond the dynamic evolution of topics. Further research is required to explore these potential influencing factors.

4.2. Relationship Topic Popularity - Toxicity

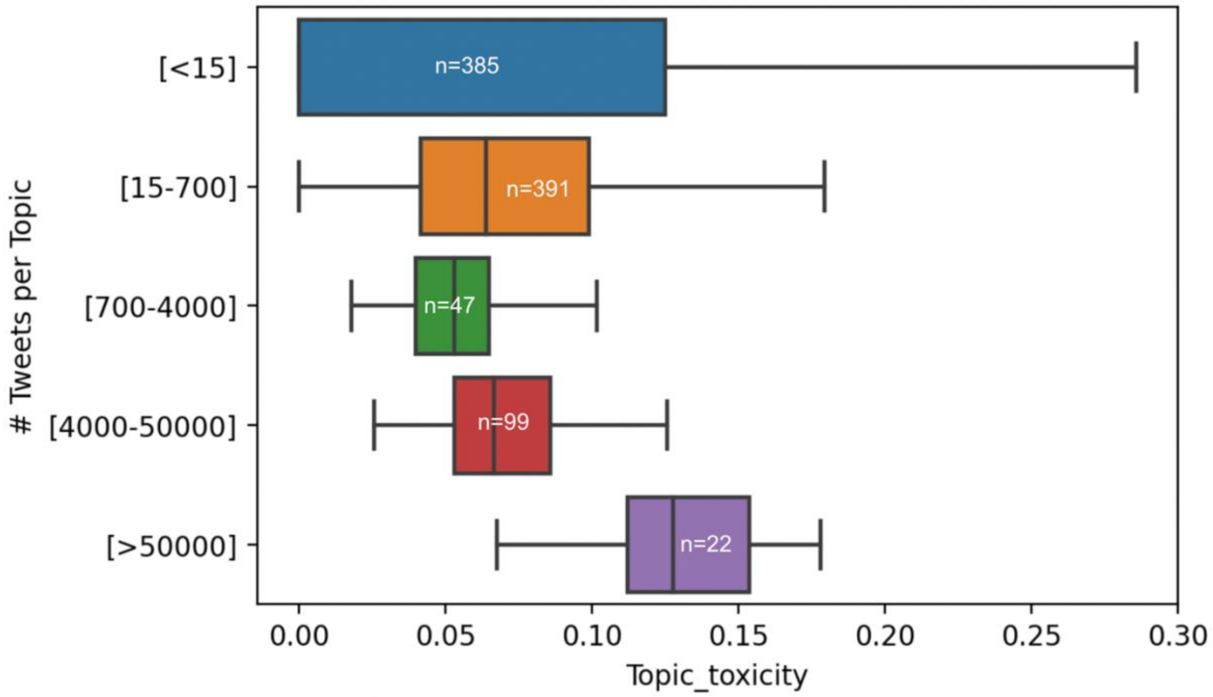

Figure 11 - Relationship between the Topic Toxicity and Topic Popularity (measured in Tweets per Topic)

Analyzing the intersection of topic popularity and toxicity offers valuable insights into the nature of conversations on social platforms like Twitter. We operationalize the notion of 'topic popularity' based on the number of tweets attributed to each topic, and 'topic toxicity' is quantified using Detoxify scores as previously discussed.

An essential question here is whether popular topics—those that garner more tweets—exhibit higher or lower levels of toxicity compared to less popular, or niche, topics.

Our findings, as demonstrated in Figures 11 and 12, suggest a parabolic relationship between topic popularity and toxicity. Specifically, if we disregard 'micro topics,' defined as those with fewer than 15 tweets, we observe a positive correlation between topic popularity and toxicity. This suggests that as a topic gains traction and becomes more popular, it also tends to become more toxic. This could be due to various factors, such as the participation of a more diverse set of users, increased intensity of debate, or even a rise in trolling or hateful behavior as the topic draws broader attention.

However, the distribution of topic popularity is right-skewed, indicating that there are a few extremely popular topics, while the majority of topics have a moderate or low level of popularity. To better understand the relationship between popularity and toxicity in this skewed distribution, we have applied a log transformation to the topic popularity variable. This transformation reduces skewness and allows for a more accurate statistical analysis of the correlation between the two variables.

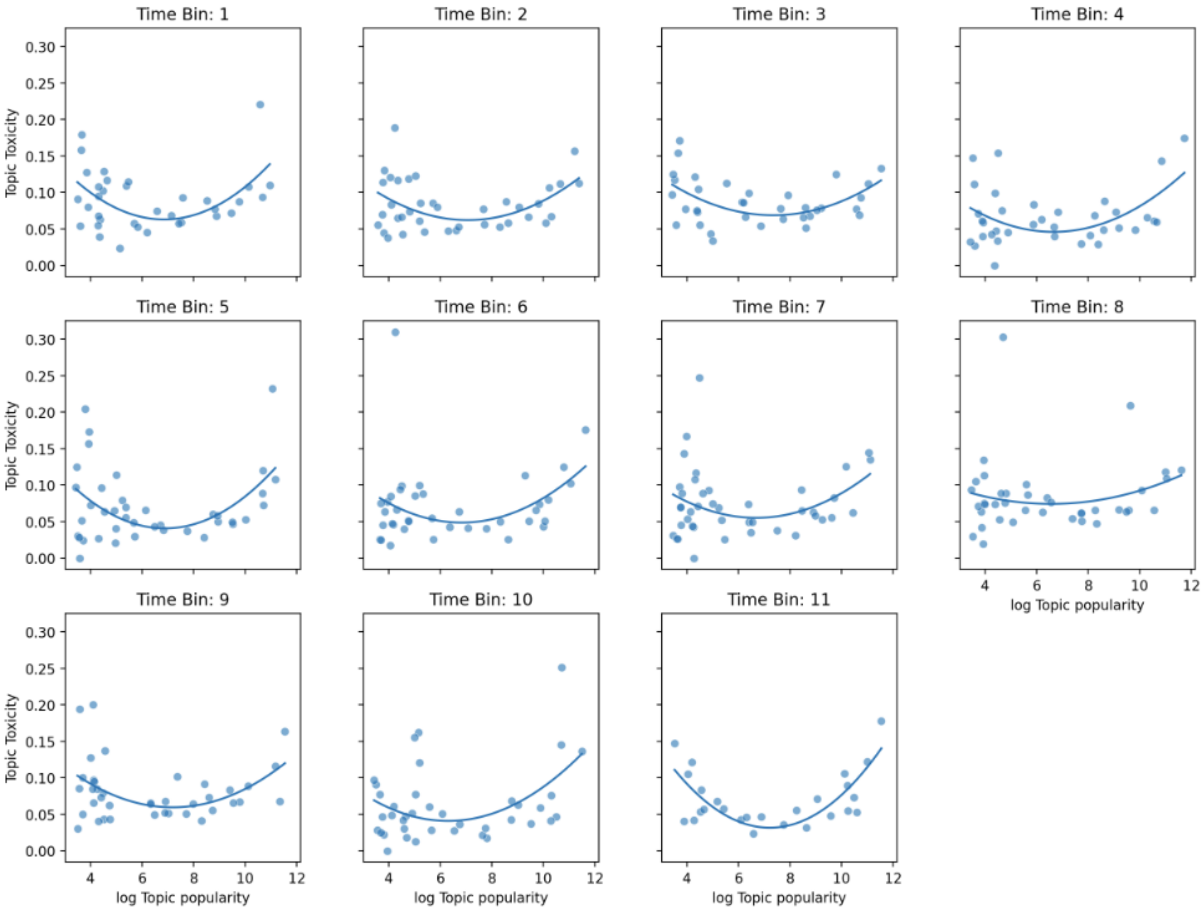

Figure 12 - Relationship between the Topic Toxicity and Topic Popularity grouped by time bin. Because of its skewed distribution the Topic Popularity was log transformed. Topics with an Popularity < 30 were discarded.

5. LIMITATIONS & FUTURE WORK

A critical aspect of our study lies in the static clustering achieved through the Generalized Scalable Dirichlet Multinomial Mixture Model (GSDMM). As we analyze the topic evolution across time bins, it becomes imperative to address the inherent limitations associated with this method.

The primary challenge originates from the inherent instability of GSDMM static clustering. Our experience has shown that even minimal variations in the input data—subtle nuances in the content of tweets—can yield significantly different clustering outcomes. This volatility contradicts our expectation of topic stagnation in instances where only insignificant variations occur between tweets from subsequent time bins. The inconsistency manifests itself as abrupt splits, merges, and other structural changes in the topics, despite the near-identical content across the time bins. The non-deterministic nature of GSDMM further exacerbates this issue, complicating the task of modeling topic evolution reliably over time.

One potential solution to this problem is the implementation of a sliding time window approach. Instead of examining the topics in discrete, non-overlapping time bins, a sliding time window would allow for continuity and greater contextual understanding, thereby offering some degree of mitigation to the observed instability. However, this proposed solution is not without its own shortcomings. Particularly, it poses a significant computational challenge, as the increase in overlapping data would inevitably lead to a corresponding increase in computation time and resource consumption.

Another alternative, inspired by the principles of k-means clustering, could prove beneficial. In this proposed "temporal k-means" approach, the initial time bin's topics would serve as the "centroids" for subsequent bins. To assign tweets to topics in the second and later time bins, we would compute the closest topic (or centroid) from the first time bin. This would form a cluster of all tweets from the second time bin linked closest to the same centroid from the first time bin. Effectively, this method circumvents the need for GSDMM beyond the first time bin, potentially providing more consistent and reliable results.

However, this solution, while conceptually sound, also introduces its own potential drawbacks. Specifically, the primary concern lies in the risk of degrading cluster performance over time. As the clustering of each subsequent bin relies on the centroids identified in the first bin, we might observe a gradual drift in topic accuracy and relevance. This could, over time, lead to a misrepresentation of genuine topic evolution, introducing a temporal bias into the model.

In conclusion, while GSDMM and its derived strategies offer valuable approaches to model topic evolution, they also possess intrinsic limitations that researchers must consider. These challenges underscore the need for continual refinement of our methodological approaches, ensuring that they evolve in tandem with the dynamism of the data we are analyzing. Future research endeavors should prioritize the development and validation of more robust, adaptive, and stable clustering techniques that better capture the subtle complexities and inherent temporal nature of evolving topics.

6. CONCLUSION

In wrapping up, our study presents an innovative methodology of dynamic Topic Evolution Modeling, that leverages an M-partite-Directed Acyclic Graph to articulate the trajectories of topics across time. Our model accommodates and showcases the dynamism inherent in Twitter conversations, allowing for a more nuanced understanding of topic evolution.

We thoroughly investigated the implications of the proximity threshold selection on the structural composition of the graph. This threshold, represented by ε, plays a critical role in how topic connections are determined - with a high ε only linking near-identical topics and a low ε causing almost all topics to link with each other. The choice of ε, therefore, has significant repercussions on the interpretability and usability of the resultant graph.

To ensure the validity of our model, we performed a two-fold validation. We conducted manual spot checks of sample topic trajectories and quantified hashtag similarity for all interconnected topics in the graph. Both these strategies robustly confirmed the consistency of our model, providing assurance of its effectiveness.

In addition to the study of topic evolution, we delved into the realm of toxicity in the discourse. Our findings drew a compelling link between topic popularity and toxicity, alluding to the notion that trending or viral topics tend to harbor more inflammatory speech characteristics compared to those less popular or niche. This finding provides crucial insights for understanding the nature of toxicity in digital platforms and can be instrumental in informing the design of interventions aimed at reducing online toxicity.

Despite the newfound understanding of toxicity in relation to popularity, our investigation into how the level of speech toxicity evolves across different transition types - namely, merge, split, emerge, disappear, stagnate - revealed a surprising trend. Contrary to what one might intuitively expect, our results demonstrated no significant difference in toxicity changes across these transition types. This is indicative of the fact that the evolution of speech toxicity remains relatively stable, irrespective of how topics transform.

In essence, our dynamic Topic Evolution Modeling approach not only provides a powerful tool for dissecting the intricacies of topic dynamics in Twitter discussions but also sheds light on the relationship between topic evolution and toxicity. By providing a detailed, data-driven lens to scrutinize the nature and trajectory of online conversations, this study makes a significant contribution to the broader research landscape of digital discourse. We hope this work can inform and inspire future research on social media dynamics, help strategize potential interventions to reduce toxicity, and promote healthier and more constructive digital spaces.

REFERENCES

Abidin, D. Z., Nurmaini, S., Malik, R. F., Rasywir, E., Pratama, Y., & others. (2019). A model of preprocessing for social media data extraction. *2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 67–72.

Abulaish, M., & Fazil, M. (2018). Modeling topic evolution in twitter: An embedding-based approach. *IEEE Access*, *6*, 64847–64857.

Alam, M. H., Ryu, W.-J., & Lee, S. (2017). Hashtag-based topic evolution in social media. *World Wide Web*, *20*, 1527–1549.

Albanese, F., & Feuerstein, E. (2021). *Improved Topic modeling in Twitter through Community Pooling*.

Araque, O., Gatti, L., & Kalimeri, K. (2020). MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-Based Systems*, *191*, 105184.

Bai, Y., Jia, S., & Chen, L. (2020). Topic evolution analysis of COVID-19 news articles. *Journal of Physics: Conference Series*, *1601*(5), 052009.

Bar-Ilan, J., & Peritz, B. C. (2009). 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.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 113–120.

Boyd-Graber, J., Hu, Y., Mimno, D., & others. (2017). Applications of topic models. *Foundations and Trends® in Information Retrieval*, *11*(2–3), 143–296.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512–515.

Derntl, M., Günnemann, N., Tillmann, A., Klamma, R., & Jarke, M. (2014). Building and exploring dynamic topic models on the web. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2012–2014.

Dey, K., Kaushik, S., Garg, K., & Shrivastava, R. (2018). *Topic Lifecycle on Social Networks: Analyzing the Effects of Semantic Continuity and Social Communities*.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1).

Gani, R., & Chalaguine, L. (2022). Feature Engineering vs BERT on Twitter Data. *ArXiv Preprint ArXiv:2210.16168*.

Garimella, V. R. K., & Weber, I. (2017). A long-term analysis of polarization on Twitter. *Eleventh International AAAI Conference on Web and Social Media*.

Geller, M., Vasconcelos, V. V., & Pinheiro, F. L. (2023). Toxicity in Evolving Twitter Topics. *International Conference on Computational Science*, 40–54.

Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G., & Plagianakos, V. P. (2018). Convolutional neural networks for toxic comment classification. *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, 1–6.

Gohr, A., Hinneburg, A., Schult, R., & Spiliopoulou, M. (2009). Topic evolution in a stream of documents. *Proceedings of the 2009 SIAM International Conference on Data Mining*, 859–870.

Grandjean, M. (2016). A social network analysis of Twitter: Mapping the digital humanities community. *Cogent Arts & Humanities*, *3*(1), 1171458.

Guo, J., Cao, L., & Gong, Z. (2021). *Recurrent Coupled Topic Modeling over Sequential Documents*.

Hanu, L. & Unitary team. (2020). *Detoxify*.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009). Detecting topic evolution in scientific literature: How can citations help? *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 957–966.

Henry, T., Banks, D., Chai, C., & Owens-Oas, D. (2018). *Modeling community structure and topics in dynamic text networks*.

HOANG, T. A., LIM, E. P., ACHANANUPARP, P., JIANG, J., & ZHU, F. (2011). On Modeling Virality of Twitter Content.(2011). *Digital Libraries: 13th International Conference on Asia-Pacific Digital Libraries, ICADL*, 24–27.

Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (2021). 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.

Hu, Y., Xu, X., & Li, L. (2016). Analyzing topic-sentiment and topic evolution over time from social media. *Knowledge Science, Engineering and Management: 9th International Conference, KSEM 2016, Passau, Germany, October 5-7, 2016, Proceedings 9*, 97–109.

Huang, F., Niranjan, U. N., Hakeem, M. U., & Anandkumar, A. (2015). *Online Tensor Methods for Learning Latent Variable Models*.

Jo, Y., Hopcroft, J. E., & Lagoze, C. (2011). The web of topics: Discovering the topology of topic evolution in a corpus. *Proceedings of the 20th International Conference on World Wide Web*, 257–266.

Johnson, K., & Goldwasser, D. (2018). Classification of moral foundations in microblog political discourse. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 720–730.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). FastText.zip: Compressing text classification models. *ArXiv Preprint ArXiv:1612.03651*.

Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, *48*(2), 354–368.

Loyal, J. D., & Chen, Y. (2020). *A Bayesian Nonparametric Latent Space Approach to Modeling Evolving Communities in Dynamic Networks*.

Masud, S., Dutta, S., Makkar, S., Jain, C., Goyal, V., Das, A., & Chakraborty, T. (2020). *Hate is the New Infodemic: A Topic-aware Modeling of Hate Speech Diffusion on Twitter*.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113.

Murayama, T., Wakamiya, S., Aramaki, E., & Kobayashi, R. (2021). Modeling the spread of fake news on Twitter. *Plos One*, *16*(4), e0250419.

Neo, S.-Y., Ran, Y., Goh, H.-K., Zheng, Y., Chua, T.-S., & Li, J. (2007). The use of topic evolution to help users browse and find answers in news video corpus. *Proceedings of the 15th ACM International Conference on Multimedia*, 198–207.

Redhu, S., Srivastava, S., Bansal, B., & Gupta, G. (2018). Sentiment analysis using text mining: A review. *International Journal on Data Science and Technology*, *4*(2), 49–53.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

Sha, H., Hasan, M. A., Mohler, G., & Brantingham, P. J. (2020). *Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. governors and cabinet executives*.

Song, M., Heo, G. E., & Kim, S. Y. (2014). Analyzing topic evolution in bioinformatics: Investigation of dynamics of the field with conference data in DBLP. *Scientometrics*, *101*, 397–428.

Stai, E., Milaiou, E., Karyotis, V., & Papavassiliou, S. (2018). Temporal dynamics of information diffusion in twitter: Modeling and experimentation. *IEEE Transactions on Computational Social Systems*, *5*(1), 256–264.

Tan, C., Lee, L., & Pang, B. (2014). The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. *ArXiv Preprint ArXiv:1405.1438*.

Udupa, A., Adarsh, K., Aravinda, A., Godihal, N. H., & Kayarvizhy, N. (2022). An Exploratory Analysis of GSDMM and BERTopic on Short Text Topic Modelling. *2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP)*, 1–9.

Viermetz, M., Skubacz, M., Ziegler, C.-N., & Seipel, D. (2008). Tracking topic evolution in news environments. *2008 10th IEEE Conference on E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services*, 215–220.

Yang, K.-C., Hui, P.-M., & Menczer, F. (2022). How Twitter data sampling biases US voter behavior characterizations. *PeerJ Computer Science*, *8*, e1025.

Yin, X., Wang, H., & Yin, P. (2018). *Agent-based opinion formation modeling in social network: A perspective of social psychology and evolutionary game theory*.

Zhang, Y., Mao, W., & Lin, J. (2017). Modeling topic evolution in social media short texts. *2017 IEEE International Conference on Big Knowledge (ICBK)*, 315–319.

Zhou, H., Yu, H., Hu, R., & Hu, J. (2017). A survey on trends of cross-media topic evolution map. *Knowledge-Based Systems*, *124*, 164–175.