2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Improving LDA Topic Modeling with Gamma and Simmelian Filtration

Evan M. Williams
Accenture
Washington, DC, USA
e.m.williams@accenturefederal.com

David Levin
Accenture
Washington, DC, USA
david.j.levin@accenturefederal.com

Ian McCulloh
Accenture
Washington, DC, USA
ian.mcculloh@accenturefederal.com

***Abstract*— Twitter has become an important tool for communication and marketing. Topic model algorithms meant to characterize the discourse of online conversations and identify relevant audiences do not perform well for this task, despite their widespread usage. This paper proposes an iterative topic model, *Gamma Filtration,* and a social network-based method, *Simmelian Filtration,* to amplify tweet-topic probability signal and reduce noise. We demonstrate the method on a novel data set collected of European Racially and Ethnically Motivated Violent Extremist (REMVE) networks on Twitter. We find that Simmelian Filtering is most successful at reducing noise as measured by perplexity. This improves our ability to detect and monitor core conversations of a community that is disseminating propaganda to increase online extremism.**

Keywords — Twitter, social media, topic modeling, latent Dirichlet allocation, LDA, social network, violent extremism

I. Introduction

Social media platforms provide people an opportunity to share thoughts and ideas, debate, and develop diverse audiences of listeners. Some use these platforms to advertise products and goods, while others may use it to promote political agendas, and still others may use it for propaganda or radicalization. As scholars study the social dynamics of online populations, a key challenge is characterizing the discourse among coherent audiences. By coherent audience, we mean a group of personas that exchange content or are likely to interact with each other's content. Characterizing their content allows scientists to investigate the underlying cultural characteristics of online communities and provides insight into understanding organic versus inorganic (bots) activity.

IEEE/ACM ASONAM 2020, December 7–10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

Perhaps the most common method of content characterization is topic modeling, specifically latent Dirichlet allocation (LDA) [1]. We find that LDA does not perform well on social media micro-blogs due to a number of factors. This paper will highlight some of the shortcomings of LDA applied to this type of social media data and propose two methods that can be used to overcome LDA limitations. The methods are demonstrated on data collected from online communities discussing European Racially and Ethnically Motivated Violent Extremist (REMVE) networks on Twitter.

This paper is organized as follows. The background provides a brief overview of LDA, summarizes related works, and highlights limitations for our application. The data section will explain the procedure used for collecting REMVE data used for demonstration. The method section will describe the two proposed data filtration methods, which we refer to as *Gamma* and *Simmelian Filtration.* The results section will compare the resulting topic models as applied to the REMVE data set. Finally, conclusions for social media analysis will be drawn.

II. Background

Latent Dirichlet Allocation (LDA) [1] is a popular form of topic modeling. It is often used for extracting latent themes from bodies of text. Even 17 years after its inception, LDA continues to be a highly influential algorithm and remains widely used in analyses of Twitter data [11][12][13]. While many researchers use LDA to explore Twitter data, LDA models using tweets are generally noisy and unstable [9]. There are several reasons why LDA may not work well with Twitter data.

Twitter data contains a high quantity of n-grams that occur once. For example, in our entire corpus, 50.0% of words only appeared once and another 15.6% only appeared twice. When documents have few co-occurring terms and many hapax legomena, LDA can fail to consistently categorize the documents well, even after thousands of iterations.

Micro-blogs like Twitter are short, meaning fewer words co-occur than in longer documents. Some documents only share one word with all other documents. This affects the consistency of term-document probability and document-topic probability in LDA.

The arbitrarily chosen initialization point, or seed, can cause LDA optimization to converge at different local optima.

Tweets generally contain only one topic per document, violating an assumption of LDA.

Some have attempted to correct LDA models for Twitter data by either aggregating tweets by author or by correcting the underlying LDA models for Twitter data [5][9]. Other researchers have proposed modifications to LDA which can improve performance on Twitter data including biterm topic modeling [14] and embedding-clustering approaches [15][16]. While we use LDA in this work, both filtering approaches we present could be used with other topic modeling approaches.

These approaches, however, generally do not assess the robustness of their models to noise or preprocessing decisions. In the authors' experience, we find LDA to be highly volatile to initialization point (seed) and to preprocessing decisions. As other researchers have found, "the inferences one draws can be extremely sensitive to the preprocessing choices the researcher makes" [2]. While topic modeling results already undergo subjective interpretation, the volatility of the underlying model suggests an additional layer of subjective decision-making.

Researchers have used community detection algorithms in conjunction with multiple runs of LDA to explore topics across discrete communities [17]. This approach requires substantially more work for the analyst, as the analyst needs to inspect multiple topics across multiple communities. Our primary contribution is the introduction of two filtration methods to improve topic coherence without repeatedly applying LDA to multiple subsets of the network.

LDA is a generative probabilistic method for topic modeling. It is common to use the measure *perplexity* to evaluate LDA models. Perplexity is used by convention in language modeling, and a lower perplexity score indicates better generalization performance [1]. To select a number of topics, *k,* one can evaluate the Perplexity Rate of Change metric proposed by Zhao [10]. For a set of documents *M*, over a set of *N* total words, perplexity can be formally defined as:

$$perplexity(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}(N_d)}\right\}$$

where w_d is the is the word vector in document *M*. Cross-fold validation is used to calculate the rate of perplexity change over candidate numbers of topics. $P_1, P_2, \ldots P_r$ denotes the average perplexities for *r* candidate number of topics. The rate of perplexity changes for topic number t_i is calculated as:

$$RPC(i) = \left|\frac{P_i - P_{i-1}}{t_i - t_{i-1}}\right|$$

To better capture model volatility in our LDA evaluation metric, we calculated 5-fold perplexity scores across 50 random seeds. We will use the mean and variance of these perplexities as a quantitative score on the effectiveness of gamma and Simmelian filtration methods.

III. Data

We collect data from an online community in Twitter to demonstrate the gamma and Simmelian filtration methods described in the next section. The authors chose a REMVE community based on background, convenience, and use this for expository purposes. We fed an expert-curated list of REMVE terms into a software from Zignal Labs, in order to capture Twitter data containing those terms posted between the beginning of August and the end of November 2019. The resulting dataset contained 633,681 tweets. Although most of

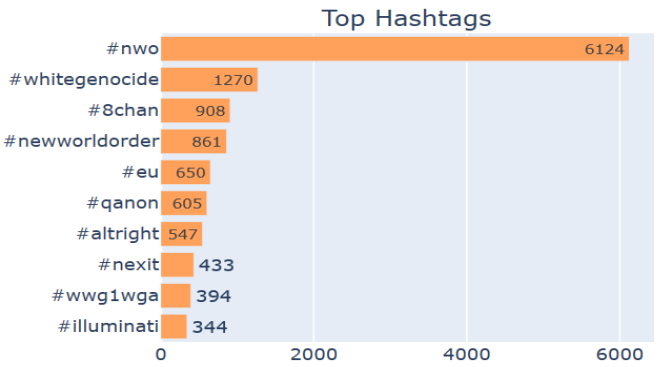

Figure 1: Top hashtags in REMVE data

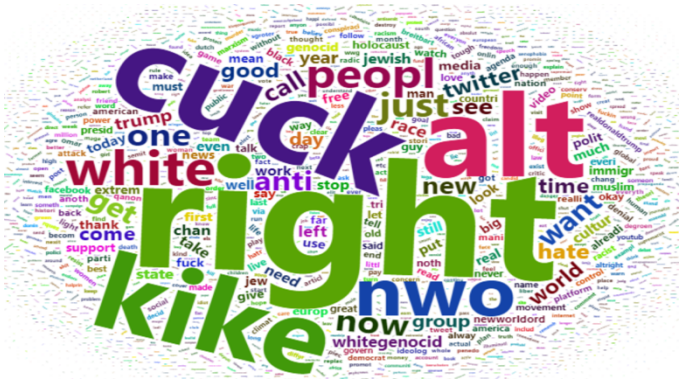

Figure 2: Wordcloud of most common words in REMVE data following English translation, stemming, lowercasing, and removal of special characters.

our data did not include user location (52.76%), the United States was by far the most commonly labeled location observed (51.66%). Labeled European accounts composed only 5.91% of the data: 37,670 tweets. To increase the size of our corpus, we trained a neural network using Keras to classify unlabeled data as "in Europe" or "not in Europe." The neural network outperformed logistic regression, decision trees, random forests, and a deeper long short-term memory neural network architecture. This increased the size of our corpus to 50,894 European Twitter documents (a.k.a tweets).

The resulting data corpus contained over 36 languages — the majority being English (28,840), Spanish (10,722), Dutch (4,563), French (3,510), and German (3,130). Because we translated these documents to English using Amazon Web Service's Translation Software Development Kit, and translation services are unavailable for Catalan, Catalan-language documents were excluded from our dataset. While this certainly results in information loss, LDA relies on n-gram co-occurrence, so we wanted tokens to be directly comparable. Figure 1 displays word clouds of both the most common pre-translated hashtags and the most common languages within our Twitter dataset. Figure 2 contains a wordcloud of our final, preprocessed corpus.

The five most common words in our preprocessed corpus (excluding stop-words) in our dataset were right (13,339), cuck (10,211), alt (9,771), kike (8,912), and nwo + newworldorder (5,755 + 820). However, it is worth noting that many mentions of the word 'kike' were made in reference to the Spanish Football player, Enrique Martínez (who goes by Kike). We were stuck by how predominantly conspiracy

theories are featured in our dataset - these will be explored in the Results section of the paper.

IV. METHOD

Topics in the REMVE data were neither stable nor semantically meaningful at this point, so we propose methods to reduce the noise and amplify the signal of meaningful topics. The first method is an iterative document-topic probability (Γ) filtering. The second is a social network analysis approach, taking the network's Simmelian Backbone [7] and running LDA on the highest-degree authors of the backboned network.

LDA models were created using the *topicmodel*s package in R [3] to return both term-document probabilities β, as well as document-topic probabilities Γ. We created LDA models with k number of topics, viewed the terms with the highest term-document scores for each topic, and read samples from the topics that seemed the least relevant. Once we confirmed a topic was capturing irrelevant information, we dropped any authors with a Γ of greater than .5 for that topic, re-filtered by term-frequency-inverse document frequency (TFIDF) at a threshold of 1, and re-ran the topic model. Put more simply, we qualitatively identify topics that seem to capture noise, any author that has a Γ of greater than 0.5 within that topic is dropped, and then we re-run TFIDF and LDA. This approach was successful in both reducing noise and amplifying the signal. However, this approach necessitates significant human oversight in the selection of irrelevant topics, and it is vulnerable to subjective decisions of the researcher. Below is one iteration of this approach `in pseudocode:`

```
X ← tokenize(X)
DTM ← DocumentTermMatrix(X)
L ← LDA(DTM)
tfidf_bound ← 1.0
drop_authors ← list()
bad_topics ← list(int(topic k to drop
after model inspection))
FUNCTION gamma_filter(L, X, bad_topics,
tfidf_bound):
        FOR (author in LDA):
                IF author's  Γ > 0.5 ∈ bad_topics:
                        drop_authors[author] ←
author
                END IF
        END FOR
        X ← X['author'] ∉ drop_authors
        X ← X ∈ X[tfidf(X['ngrams']) >
tfidf_bound]
        Return X
```

We refer to this method as Gamma Filtering.

Our second approach took advantage of the network structure of Twitter data. A social network can be represented as a graph G where vertices V represent users, and the edges E represent ties between them. In this case, V are all authors and twitter handles that were present within tweets and E are symmetrized valued sums of all author to Twitter-handle interactions (retweets, mentions, replies).

We propose using community detection algorithms to identify core communities within a network as a filtration step prior to running LDA on social media datasets. We refer to this approach as Simmelian Filtration. We used the non-parametric variant of the Simmelian Backbone algorithm proposed by Nick et al. [7] to reduce our dense Twitter social network down to its most densely connected triads. For each node, the algorithm ranks neighbors by tie strength, determines redundancy of small ties, and filters ties that are not week or redundant. We chose this algorithm due to its strong roots in social science theory, but it is likely that other community detection algorithms could successfully reduce noise in social media datasets prior to running LDA. Simmelian bonds are triangles, where each corner is a user and where each edge is a bi-directional bond. Nick et al. [7] outline the social science theory:

"A group's capabilities (and constraints) are then determined by two dimensions: the redundancy of internal edges and the non-redundancy of external edges (along the vertical axis). Burt argues that a group achieves maximum performance when its internal edges are strong, and its external edges are weak: 'while brokerage across structural holes is the source of added value, closure can be critical to realizing the value buried in the structural holes.'" [7]

This approach has three primary benefits. First, as Simmelian groups are highly connected triads, groups identified by a Simmelian Backbone are more likely to exhibit strong homophily, or a tendency to interact with similar individuals. It has previously been found that homophily in Twitter produces "a tendency of users to produce like-minded information, [where] individuals are disproportionately exposed to like-minded information" [4]. The intuition is that densely connected users should be sharing similar content, which in turn exhibits a strong signal representative of a larger conversation.

Second, Simmelian groups are likely to have a more limited repertoire of topics. Within self-categorization theory, depersonalization describes a process of self-stereotyping. This is where, within a relevant social category, "people come to see themselves more as the interchangeable exemplars of a social category than as unique personalities defined by their differences from others" [8]. That is to say, individuals base their beliefs, goals, and behavior as members of the group. In theory, this should elucidate clearer themes and terminologies.

Third, despite a dataset dominated by English, densely connected triads naturally form in different languages. This amplifies conversations and communities that form in non-English languages.

These qualities reduce the number of likely topics discussed and the range of terms used to describe the topics. In other words, down-selecting on Simmelian Backbones should increase central tendencies and reduce variance. This should serve to clarify signal strength of any clustering that is to be detected by LDA or other NLP topic-modeling procedures.

The performance improvement of the filtering methods are measured by comparing the perplexity scores of the different methods. A lower perplexity score indicates a clearer signal to noise ratio and a more successful filtration method.

TABLE I. 5-FOLD PERPLEXITY OVER 50 SEEDS (LOWER IS BETTER)

Method	*Mean*	*Variance*
LDA on full tweet Corpus	1,348.3	325.6
Iterative Gamma Filtration	936.4	2,456.3
Simmelian Filtration	257.6	21.8

V. Results and Analysis

Traditional LDA was highly volatile with our data. As shown in Table 1, over 50 different random initialization points, the mean perplexity was 1,348.33 with a variance of 325.58. This is partially because the initialization parameters change "the approximate optimizing solutions to LDA may converge to a different local optimal point for the same dataset" [10]. This problem was not ameliorated by using Correlational Topic Models, Variational Expectation Models, Fixed Variational Expectation Models, or Gibbs Sampling. The models were volatile to topic number, to initialization point (seed), and to cleaning and preprocessing decisions.

We then ran 7 iterations of our Gamma Filtration Approach. This reduced our corpus to 7,874 tweets. While the top terms seemed qualitatively meaningful, this approach was not robust over changing seeds. This approach yielded a variance of 2,456.3.

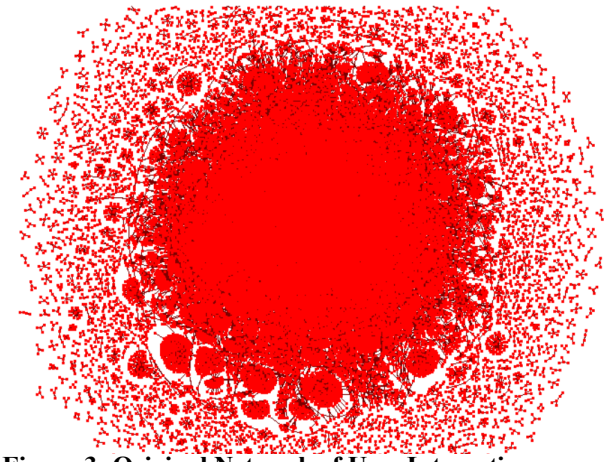

Figure 3: Original Network of User Interactions

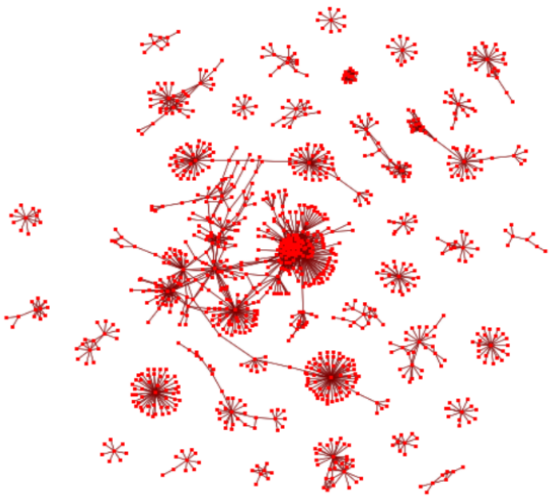

Figure 4: Simmelian Backbone of User Interactions

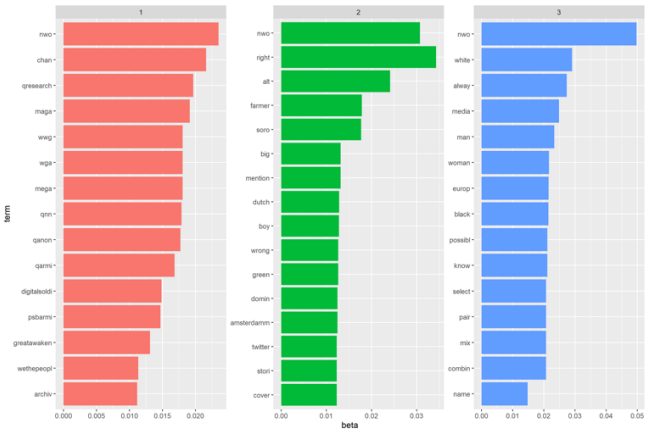

Figure 5: Final Topic Model Showing Distinct Topics

Finally, we ran the Simmelian Backbone algorithm on our dataset. We then took the 30 authors with highest degree, and verified that each were in our target population. Twenty-five of these 30 authors were relevant and were disseminating REMVE content. The remaining five authors were dropped for either being American or anti-fascist.

This technique was highly successful at clarifying signal strength, and at revealing themes of core conversations. However, after reducing our final LDA model to only 933 documents (Tweet-level), top terms were still volatile to changes in initialization point. Despite this limitation, the distribution of top topics across sequential *k* for a given seed was qualitatively more stable. Simmelian Filtration performed significantly better than Traditional LDA and Gamma Filtration, yielding an average perplexity of 257.6 and a variance of 21.8.

The Simmelian backbone is also useful for further analysis of the social network of agents in the online discussion. The Simmelian filter reduces complexity in the network. Figure 3 shows the original network of user interactions. Figure 4 shows the backbone network, revealing a central core network with satellite clusters.

Our final model found high-degree, densely connected triads in English, Spanish, German, Dutch and French. Figure 5 shows topics from the final model that shows distinct topics.

In a semantic hashtag network run over our full dataset (where each pair of Vertices V is a co-occurring hashtag within a Tweet), #nwo (new world order) displayed the highest centrality by most common centrality metrics [6]. It had the highest degree centrality, meaning that it was used in tandem with more hashtags than any other hashtag in our dataset. #nwo also had, by far, the highest betweenness centrality of any hashtag, followed by #newworldorder, with respective scaled betweenness centralities of 0.436 and 0.104. This suggests that the "new world order" conspiracy serves as a primary bridge for REMVE activists across European countries. The conspiracy theories denoted by #whitegenocide and #qanon were also highly central hashtags within our network.

VI. Conclusion

This paper presents a method for reducing noise and better identifying topics in online Twitter discourse. The application of Gamma and Simmelian filtering reduces the perplexity score of resulting LDA topic models and qualitatively results

in more distinct topics. An extensive exploration of this method across different datasets is beyond the scope of this paper. We leave that for future research. Still, this method introduces a novel approach for improving LDA performance in social media data.

There are several limitations to this study. This was focused on a single issue, REMVE content. Future work should explore other content areas and possibly focus on the relationships between areas. This study also focused on two methods of complexity reduction. A comparison of additional complexity reductions may prove successful. The approach laid out in this paper to use perplexity as a measure of success would be useful in a future comparison on approaches.

Throughout this analysis, we were struck by how, at every level, conspiracy theories united disparate REMVE groups. There were many different centrality measures for networks, and the "new world order" conspiracy theory dominated many of them. These findings suggest that conspiracy theories are playing a significant role in uniting disparate REMVE communities across Europe. Conspiracy theories are serving as an international bridge between REMVE movements. Combatting REMVE may necessitate additional attention to the online communities where conspiracy theories incubate and spread. More research will need to be done to understand the overlap, interaction, and symbiosis of conspiracy theories and REMVE. Additionally, policymakers and intelligence communities may need to broaden the scope of what is considered REMVE, in order to effectively combat its spread.

With the steady increase of social media use, it is important to develop better methods to identify relevant discourse online and contrast signals of relevant content from noise. Micro-blogs, such as Twitter, present unique challenges for natural language processing due to the shortened length of documents, slang, and other issues. This increases the need and importance of improved methods for content analysis. This paper contributes to the literature in presenting a method of signal to noise amplification as well as an approach to compare different methods for improved performance in a social media/micro-blog context.

REFERENCES

[1] Blei, D. M., Ng, A. Y. Jordan, M. L. (2003). Latent dirichlet allocation. *Journal of machine Learning research* 3.Jan (2003): 993-1022.

[2] Denny, M. J., Spirling A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis* 26.2 (2018): 168-189.

[3] Grün B, Hornik K (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13): 1–30. doi: 10.18637/jss.v040.i13.

[4] Halberstam, Y., Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of Public Economics* 143 (2016): 73-88.

[5] Lim, K. W., Chen, C., & Buntine, W. (2016). Twitter-network topic model: A full Bayesian treatment for social network and text modeling. arXiv preprint arXiv:1609.06791.

[6] McCulloh, I., Armstrong, H., Johnson, A.N. (2013) *Social Network Analysis with Applications*. Hoboken, NJ: Wiley

[7] Nick, B., Lee, C., Cunningham, P., & Brandes, U. (2013). Simmelian backbones: Amplifying hidden homophily in facebook networks. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2013.

[8] Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Basil Blackwell, 1987.

[9] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. *European conference on information retrieval*. (pp. 338-349). Springer, Berlin, Heidelberg.

[10] Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. In BMC bioinformatics (Vol. 16, No. 13, p. S8). BioMed Central.

[11] Yeung, Neil, Jonathan Lai, and Jiebo Luo. "Face Off: Polarized Public Opinions on Personal Face Mask Usage during the COVID-19 Pandemic." arXiv preprint arXiv:2011.00336 (2020).

[12] Feng, Yunhe, and Wenjun Zhou. "Is working from home the new norm? an observational study based on a large geo-tagged covid-19 twitter dataset." arXiv preprint arXiv:2006.08581 (2020).

[13] Xue, Jia, et al. "Twitter discussions and emotions about COVID-19 pandemic: a machine learning approach." Journal of medical Internet research.

[14] Cheng, Xueqi, et al. "Btm: Topic modeling over short texts." IEEE Transactions on Knowledge and Data Engineering 26.12 (2014): 2928-2941.

[15] Jónsson, Elıas, and Jake Stolee. "An evaluation of topic modelling techniques for twitter." (2015).

[16] Angelov, Dimo. "Top2Vec: Distributed Representations of Topics." arXiv preprint arXiv:2008.09470 (2020).

[17] Surian, Didi, et al. "Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection." Journal of medical Internet research 18.8 (2016): e232.