Topic Flow Modeling with Latent Communities for Microblog Conversations

Chad Crawford, Sandip Sen
The University of Tulsa
chad-crawford@utulsa.edu, sandip-sen@utulsa.edu

Abstract

With the sudden and significant growth of participation in social media, there has been a shift in need for intelligent agents to accommodate its users. One of the most salient aspects of social media is its conversational aspect, where users can interact to share and evolve their opinions and beliefs. The conversational context of social media posts can influence their topics, and can often be more influential than the community characteristics and features. The nature of how topics may evolve in a conversation, or the "topic flow" of a conversation, is strongly dependent on the personality of users that participate in the conversation. We build on prior work on topic models and introduce a novel, unsupervised statistical model of topic flow on social networks, the Latent Category Topic Flow Model (LCTFM). LCTFM learns to identify topics from a set of conversations where documents can have their own conversational trends. For example, the topic of the response to a post may be influenced by the setting of the conversation or the personality of the participants. LCTFM significantly outperforms other topic models that both ignore and account for conversational information.

1 Introduction

Social media has become one of the primary sensory and actuatory tools with which today's citizens interact with their social world. People rely on social media for news and information as well as notifications about life events of others in their social network. They routinely share their personal information on a variety of social media platforms supporting communication modalities ranging from tweets to blogs and from pictures to posts. They readily express their worldview and opinions about sociocultural issues and topics ranging from politics to the environment and from scientific arguments to religious debates in a multitude of forms. Users not only upload isolated items, but often engage in repeated and sustained interaction with other users. While some of these interactions are informal or can be termed friendly exchanges, other interactions are more substantive and reflect the user's opinions and beliefs on current issues and topics of importance to them. The latter often involves repeated communication in response to those posted by other users and can be best understood in their conversational context. Whether we want to understand the belief structure of individual users, for better assisting them to express themselves or to retrieve information of value to them, or develop a more comprehensive picture of the emerging and shifting landscape of public opinion on turnkey social issues and trending topics, which can be of much value from policy makers to marketing professionals, the ability to analyze interaction patterns on social media to identify conversational topics, their relative importance and their mutual influence is a critical functionality that will enable us to develop a more comprehensive understanding of why and how users interact, what topics are of primary concern to them and how they influence or are influenced by social media conversations on those topics.

One of the noteworthy aspects of social media activities is its conversational nature. Web forums and websites such as Reddit and Twitter are dynamic platforms in which users discuss a multitude of topics. Social media posts[1] are often short; on Twitter, for example, tweets are limited to 280 characters. In some cases, a post might not have any words at all, containing just a link to an image or news article. Social media is distinct from other forms of document corpora in several ways: Posts (or documents) are conversational and usually contain language and terminology that is comprehensible only to some sub-community. In addition, many social media outlets such as Twitter, Facebook, YouTube and Reddit support tree-structured conversations, in which a single post may have multiple, distinct responses. The "conversational context" of each post has a direct and formative impact on its topic and is usually even more influential than the demographic and other characteristics of the community.

There has been an explosion of user activity on social media in recent years and the associated preponder-

[1] We use the term *post* to refer to text-based user communications on social media platforms.

ance of accessible datasets have piqued interest among researchers in analyzing and experimenting with such datasets. Some of these collections contain documents which are typically short, can benefit from topic extraction for in-depth analysis, and is informative only when placed in the associated context. The most significant factors that help determine the topics in a document is the context of a conversation. We define the **conversational context** to be a situation in which the author of a reply decides the topic(s) for their response by responding independently to topic(s) of the parent document. The focus of this work is on the concept of **topic flow** which characterizes how topics evolve from parent to response documents. We use the following hypothesis about the nature of topic flow on a social network:

Assumption 1 *The topic (or topics) of a response document to a source document are determined by the topics represented in the source document.*

However, the topic flow may not be constant and can vary from post to post. Topic flow patterns can change based on the personalities of authors or the context in which the conversation takes place. For example, individuals on polar opposites of some ideological spectrum will respond differently to topics on key controversial issues. Similarly, if a conversation begins on one topic, it is more likely to return to that topic much later in the discussion. In both cases, a prediction about the topic of a response is influenced by meta-information outside of the text of these documents. We introduce another hypothesis to address such scenarios:

Assumption 2 *Topic flow patterns can be influenced by external factors.*

This work focuses on identifying unique topic flow characteristics that can be identified through meta-information. We present the Latent Community Topic Flow Model (LCTFM), a general method for identifying latent communities that have distinct but predictable topic flow characteristics. The additional information provided to the LCTFM is a single group that each document belongs to – this may be author information or the conversation that the document belongs in. To verify that these assumptions are true for social media conversations, the LCTFM is compared to several competing algorithms that have been developed for social media and/or conversational texts. We find that the LCTFM has better predicative capabilities compared to existing topic models, and finds topics that are human-interpretable.

2 Related Works

Understanding human dialogue has been an important challenge in artificial intelligence (AI) for problems such as negotiation [Gutnik and Kaminka, 2004] or mediation [Barlier *et al.*, 2016]. There has been a large number of works on constructing conversational agents, that can have conversations with humans and be indistinguishable from other humans [Barlier *et al.*, 2016], or to serve further purposes such as establishing trust [Bickmore and Cassell, 2001].

There are several basic topic models with multiple topics per document – Probabilistic Latent Semantic Analysis [Hofmann, 1999] (PLSA) and Latent Dirichlet Allocation [Blei *et al.*, 2003] (LDA). The primary difference between PLSA and LDA is that topic distributions among documents and word distributions among topics share a common set of prior information. LDA generally outperforms PLSA on topic modeling measures because the Bayesian priors on the topic distributions make it less susceptible to overfitting topic distributions for new documents.

The limitation of LDA is that it assumes independence between documents. PLSA and LDA are especially weak for short texts due to high sparsity. Subsequent works on topic modeling enrich the topic distributions by incorporating information from a variety of link types. One rich source of inter-document relationships can come from incorporating authorship information. Such methods effectively aggregate documents with a single author, and then treat documents with multiple authors (such as scientific papers) as some combination of each author's topic distribution [Steyvers *et al.*, 2004; Rosen-Zvi *et al.*, 2010].

Hyperlinks among webpages or citations in scientific papers represent links between documents that may imply some degree of similarity between them. Such links can be generalized to any predefined relationships [Daumé, 2009]. The hypothesis behind linked documents is that they will share topics along some dimension, and the corresponding topic similarities can be used to enhance web search or find relevant scientific papers with some topic. Generally, these models extend LDA so that linked documents have a weak influence towards being similar along some dimension [Sun and Gao, 2008; Nallapati *et al.*, 2011]. A weakness of hyperlinked topic models, addressed by [Liu *et al.*, 2009], was that the vast majority of documents connected through a citation were strongly dissimilar. To address this issue, the Topic-Link LDA model [Liu *et al.*, 2009] simultaneously categorizes authors into latent social communities that can explain citation links that can be explained by topic similarity vs. links that can be explained by community similarity (for example, citing papers due to authors attending similar conferences or the same school). The assumption behind hyperlinked models is that document links imply *content similarity*, which is not necessarily compatible with conversational links. In conversations, response topics are not necessarily similar, but may still be predictable.

Temporal links are another significant influence on topic. Such links have been one of the primary focuses of topic analysis on social media, since predicting future trends is an important problem for businesses. Dynamic topic models, which allow changes in the prior topic distribution over time, have been developed for discrete [Blei and Lafferty, 2006] and continuous [Wang *et al.*, 2012a] scenarios. In the Dynamic Topic Model (DTM) [Blei and Lafferty, 2006], the Dirichlet hyperparameters to the topic and word distributions are assumed to undergo some Gaussian noise. Eventually, the model hyperparam-

eters will undergo significant changes as a community changes its interest in topics or the topics themselves change. There are other discrete-time topic models, such as TM-LDA [Wang *et al.*, 2012b] which use a similar topic-flow style assumption. However, TM-LDA is a non-generative application of LDA to learn and predict topic transition trends over time. We are interested in developing a generative model, in which conversational context can help identify latent topics that would be hard to discover under the regular independence assumption. There has also been some work on recognizing dialogue acts in text using topic models, such as in [Ritter *et al.*, 2010b]. These are based on conversational text, but focus on topics changing based on the dialogue acts of a conversation (such as, a question and a response) rather than topic transitions dependent on those topics (such as a topic evolving from sports to good bars to visit in some area).

3 Models

This section first presents the basic Markov Topic Flow Model and then the Latent Category Topic Flow Model, which is an extension of the former.

3.1 Markov Topic Flow Model

The first model we develop is an extension of Ritter et al.'s Conversation model in [Ritter *et al.*, 2010a]. Topic flow transitions are modeled using a Markov chain specified by a $K \times K$ matrix A, where $[A]_{kj}$ represents the probability of a response having topic j for a parent document with topic k. Unlike LDA or PLSA, the MTFM assigns a single topic per document. Since several social media posts are typically very short, there is often not enough space to express a complex combination of topics. This model is structurally similar to a Hidden Markov Model for topics, and is an extension of Ritter's conversation model [Ritter *et al.*, 2010a] to tree-structured conversations.

Each root document r draws a topic $z_r \sim \text{Categorical}(\pi)$, where π is a $K \times 1$ vector of topic probabilities, and z_r is a one-hot vector where $z_{rk} = 1$ where r has topic k. Any subsequent document d, with parent document p, draws a topic $z_d \sim \text{Categorical}(A^\intercal z_p)$. For each document d, each word of the document is generated from $\text{Categorical}(\theta_{z_d})$, where θ_{z_d} is an $M \times 1$ vector of word probabilities for the z_dth topic.

Parameter estimation can be performed by the Expectation Maximization (EM) algorithm. Since the model is tree-structured, inference is exact and the marginal latent variable distributions can be calculated using Pearl's Belief Propagation algorithm [Neapolitan and Others, 2004]. The optimal model parameters are similar to the Hidden Markov Model:

$$\pi_k = \frac{\sum_{s \in \mathcal{S}} Q_{sk}}{||\mathcal{S}||}, A \propto Q^T B Q, \theta \propto C^T Q.$$

Where Q is the $N \times K$ latent variable probability matrix, $\mathcal{S}$ is the set of "source" documents (i.e. documents that start a conversation), B is an $N \times N$ adjacency matrix where $B_{ij} = 1$ if the jth document is a response to document i, and C is an $N \times M$ matrix of word counts per document. B and C are sparse matrices, so these large matrix products can be reduced into much faster operations on the nonzero elements of the matrices.

3.2 Latent Category Topic Flow Model

Characterizing all conversations with a single transition matrix is limiting, and there may be scenarios in which latent contextual factors can influence topic flow. For example, suppose one user has a tendency to discuss only topics they are interested in; such conversations with these users may start from any point but would be expected to eventually gravitate towards their topics of interest. The starting point of a conversation may also influence how topics evolve. Discussions about divisive political issues will evolve differently in a debate-driven community compared to a community whose members agree on the issues. We are interested in methods in which arbitrary grouping of documents (by author, conversation, etc.) have distinct topic flow characteristics.

Such groups often contain a small number of documents. Conversations can be short, and the majority of social media users are not heavily active in conversing with others. There are $K(K+1)$ parameters per group and hence a total of $GK(K+1)$ parameters. A significantly higher number of posts would be needed to learn these parameters. To resolve this issue, groups are probabilistically assigned to latent communities, each with their own topic flow characteristics. These latent communities can be learned simultaneously with the topics in a joint Latent Community Topic Flow Model (LCTFM).

Suppose that, in addition to K topics, there are G groups and C latent communities. Each community $c = 1, \ldots, C$ has an associated topic transition matrix $A^{(c)}$ and initial topic matrix $\pi^{(c)}$. Group-community relationships are represented by the latent variables $Y = \{Y_1, \ldots, Y_G\}$. Each y_g is a $C \times 1$ one-hot vector where $y_{gc} = 1$ if and only if the group g belongs to community c. With the addition of communities, this model has $CK(K+1) + KM$ parameters, which is substantially more than the MTFM but small in comparison to the size of even modest social media corpora.

Parameter Estimation

The addition of category links between documents leads to the possibility of the underlying probabilistic network violating the polytree property. Inference on such networks is generally NP-complete and requires approximation methods [Neapolitan and Others, 2004]. To resolve this issue, we construct a variational distribution to approximate the true latent variable posterior distribution.

Variational inference is preferred over other sampling-based methods since each subproblem can be solved efficiently. Let Z be the set of random topic variables for all documents, and Y be the set of all random group variables for all authors. Then $p(Z, Y|\mathcal{D})$ is the desired conditional distribution on both topics and groups that would be necessary for performing the E-step of the EM algorithm.

Instead, we assume that the joint latent distribution $q(Z, Y)$ can be factorized as $q(Z, Y) = q_1(Z)q_2(Y)$, and then update the distributions using the following iterative procedure [Blei *et al.*, 2017]:

$$q_1^{(n+1)}(Z) = \exp\left(\mathbb{E}_{q_2^{(n)}}[\log P(Z|Y, \mathcal{D})]\right), \quad (1)$$

$$q_2^{(n+1)}(Y) = \exp\left(\mathbb{E}_{q_1^{(n)}}[\log P(Y|Z, \mathcal{D})]\right). \quad (2)$$

With Bayes' theorem, we expand the posterior distribution on q_2 to a tractable term:

$$\log q_2(y_{gc}) \propto P(Z, \mathcal{D}|y_{gc})P(y_{gc}) = \log \phi_c + \sum_{s\in\mathcal{S}_g}\sum_k q(z_{sk}) \log P(z_{sk}, \mathcal{D}|y_{gc}) + \sum_{p,d\in\mathcal{R}_g}\sum_k\sum_{k'} q(z_{pk})q(z_{dk'}) \log P(z_{pk}, z_{dk'}, \mathcal{D}|y_{gc}),$$

where $\mathcal{S}_g$ are the set of root documents belonging to group g and $\mathcal{R}_g$ are the parent-reply document pairs.

The update of q_1 can be performed via two methods. The first of these is to apply the coordinate ascent procedure to each topic distribution of the documents:

$$\log q_1(z_{dk}) \propto \log P(w_d|z_{dk}) + \sum_{k'} q(z_{pk'}) \sum_c q(y_{g_dc}) \log P(z_{dk}|z_{pk'}, y_{g_dc}) + \sum_r\sum_{k'} q(z_{rk'}) \sum_c q(y_{g_rc}) \log P(z_{rk'}|z_{dk}, y_{g_rc}).$$

This method requires multiple updates of each topic distribution to adequately address the parent-reply dependencies between documents, which can be relatively slow. If there are few groups per conversation, there is another optimization which can significantly improve performance. When Y is known, $P(Z|y, \mathcal{D})$ is tractable to compute using Pearl's Belief Propagation algorithm. Suppose G groups participate in a conversation. Then computing the expectation term $\sum_y q(y) \log P(Z|y, \mathcal{D})$ requires computing C^G different conditional distributions. This can be significantly faster than variational updates since the conditional distributions can be computed independently. If each conversation is its own group, then $G = 1$ and only C computations are necessary for an exact update along all topic dimensions.

Since the topic parameters are sensitive to both the topics of their parent or any replies as well as any document groups, our procedure will focus on topic convergence for each conversation before updating the group-community parameter Y. This procedure is significantly slower than the update procedure for the Mixture of Unigrams or MTFM models, due to the fact that both variational distributions must be updated repeatedly until convergence.

The maximization step of the EM algorithm is similar to the MTFM. Let $U_{ik} = q_1(z_{ik})$ and $V_{ik}^{(c)} = q_1(z_{ik})q_2(y_{gc})$. Then, the optimal values for the model parameters are:

$$\phi_g \propto \sum_u q_2(Y_u = g), \pi_k^{(g)} \propto \sum_{i\in\mathcal{S}} V_{ik}^{(g)},$$
$$A^{(g)} \propto U^\intercal B(V^{(g)})^\intercal, \theta \propto U^\intercal C.$$

Similar to the MTFM, the matrix updates of $A^{(g)}$ and θ are fast since B and C are very sparse.

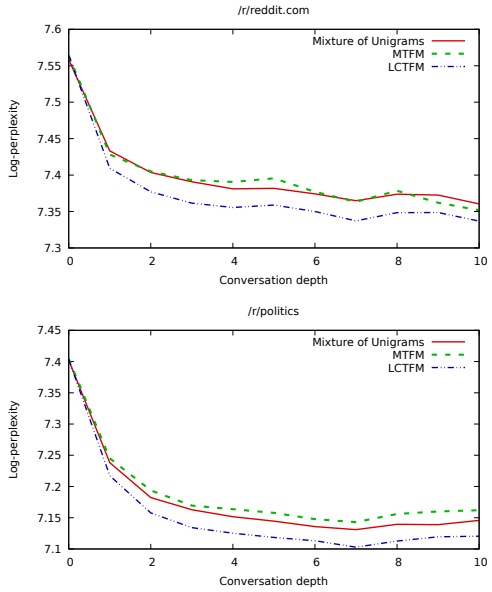


Figure 1: Comparison of conversational log-perplexity values by conversation depth for the Mixture of Unigrams, MTFM and LCTFM. Each result is averaged over 10 trained models. (top) Log-perplexities for the `/r/reddit.com` corpus. (bottom) Log-perplexities for the `/r/politics` corpus.

4 Experimental Results

4.1 Dataset

The corpora used for evaluation was scraped from the social media website Reddit. Reddit is a link-sharing website organized into subcommunities, called "subreddits", that focus on particular topics such as news, sports or politics. Users can post *submissions* to subreddits, which are generally links to articles, images or videos. Comments can be posted on submissions, which are generally short remarks made by users with regards to the subject of the submission. Reddit also allows users to respond to comments within submissions, which can lead to tree-structured conversations.

This work uses two manually-collected corpora – one from a US politics-oriented subreddit `/r/politics`, and another general discussion subreddit `/r/reddit.com`, which was made inactive in 2013. Both subreddits have distinct dynamics. Discussion on `/r/politics` revolves around ongoing events, while discussions `/r/reddit.com` can span a wide number of subjects. We collected the

most popular posts between January and February from `/r/politics`, and the most popular posts of all time from `/r/reddit.com`. We found that grouping documents by conversation was more effective in analyzing the Reddit datasets over authorship information, and thus the conversation grouping method is used for evaluation.

4.2 Preprocessing

The model was susceptible to noisy words and documents, so these were filtered out before training the topic flow models. We filtered out any conversations that were only one level deep – that is, conversations consisting of root document responses without any subsequent responses. There were several conversations that constituted a huge portion of the corpus. Conversations with more than 100 participating documents were filtered out to prevent them from overwhelming the rest of the corpus.

Unlike Twitter, the Reddit communities do not generally have a strong in-community vernacular that would make tokenization or parsing challenging. The texts were filtered for Reddit-specific formatting tags such as links or emphasis. Noisy words were filtered out using a standard stopword removal procedure as well as filtering using Part-of-Speech word tagging. Each of the document tokens are tagged using the Stanford POS tagger [Toutanova *et al.*, 2003]. The Stanford POS tagger tags each token in the document with a part-of-speech tag (such as noun, verb, adjective, etc). We manually filtered out noninformative tags such as prepositions, symbols, the word `to`, etc.

Finally, the resulting words were stemmed to decrease the sparsity of the data. Uncommon words (less that 20 appearances overall) and common words (appears in more than 20% of documents) were also filtered. The final `/r/politics` dataset contains 61966 documents, 3991 words and 5025 conversations. Each document had 15.3 words on average, and each conversation had an average depth of 5.34 responses and an average of 12.33 documents total. Note that this will be especially hard for regular state-of-the-art models such as LDA to identify, since the data remains very sparse.

4.3 Perplexity evaluation

Unlike regular unsupervised models, perplexity evaluation on conversational models generally requires that documents in the middle of a conversation are not left out during training, as it would require the model to marginalize

We define the *test-train log-perplexity* to be the average perplexity per word in a conversation, where each conversation belongs, in its entirety, to either the training or the testing set. For a set of documents $\mathcal{X} = \{X_1, \ldots, X_n\}$ where X_i is a set of conversations, the log-perplexity of a document is [Blei *et al.*, 2003]

$$perplexity(\mathcal{X}) = \exp\left(-\frac{\sum_i \log P(X_i)}{\sum_{d\in\mathcal{X}} M_d}\right)$$

where M_d is the number of words in document d. In order to test a model's ability to predict the "direction" of a conversation, the *conversational log-perplexity* splits conversations by depth; documents that are fewer than h responses from a conversation root are included in the training set $\mathcal{X}_h$, and those greater than h are placed in the test set $\mathcal{X}'_h$. We may then use the conditional probability $P(\mathcal{X}'_h|\mathcal{X}_h) = \frac{P(\mathcal{X}'_h,\mathcal{X}_h)}{P(\mathcal{X}_h)}$ in place of the joint probability term $P(X_i)$. We use the test-train log-perplexity for model selection and the conversational log-perplexity for comparing models.

Marginal likelihood calculation is challenging for both the MTFM and LCTFM, as it requires integrating over all latent variable combinations. Continuous topic distribution models like LDA construct a lower-bound on the likelihood function, referred to as the *evidence lower bound*, which can be used as a lower-bound approximation on the marginal likelihood. For the MTFM and LCTFM, we use an importance sampling method [Wallach *et al.*, 2009] to estimate likelihoods. Empirically, importance sampling was efficient and had a fast convergence rate.

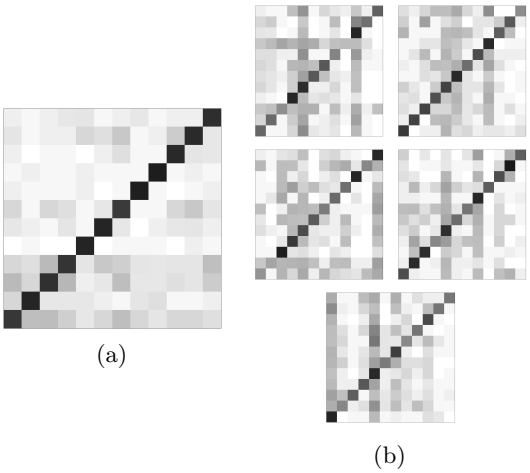
Under the train/test log-perplexity metric, the MTFM performed best with approximately 12 topics. The LCTFM performed best with 5 user groups and 12 topics. In comparison to other corpora, these models had a relatively small number of topics. However, since the corpora were fairly small (fewer than 100,000 documents) with short documents and was mostly relevant to one subject, few topics were needed to describe most trends.

The conversational log-perplexities of the Mixture of Unigrams, MTFM and LCTFM are compared in Figure 1. A lower log-perplexity value indicates better performance, so the LCTFM is the best algorithm by a significant margin. Note that unlike increasing the dimensionality of a probabilistic model, the log-perplexity is not a monotonically decreasing function of depth, since unpredictable document occurring late within a conversation can increase log-perplexity. LDA was left out of this comparison because it significantly underperformed the other models to the point that it was hard to distinguish the differences between these three models. Surprisingly, the MTFM and Mixture of Unigrams were fairly competitive for both approaches, and yet the LCTFM makes significant improvements as it is provided more conversational links.

Another interesting question is if the identified topics for this model were coherent. "Coherence" is defined as how easily a topic can be summarized by a human. There are several measures of topic coherence that have been found to correlate with human judgments of coherence. For our work, we will use the *UMass* measure [Mimno *et al.*, 2011], which is a function of co-document frequencies among the top words for each topic that has been found to agree with human judgment. The coherences are calculated using the software package Palmetto [Röder *et al.*, 2015] with a Wikipedia corpus for the co-document frequencies. The results are shown in Table 1, for topics trained on the `/r/reddit.com` corpus. We preferred to use the `/r/reddit.com` corpus because many of the posts are dated before the Wikipedia corpus for the UMass measure was collected. This is not true for the `/r/politics`

	Mixture of Unigrams	MTFM	LCTFM	Biterm Topic Model
Score	-1.73 ± 0.20	$\mathbf{-1.64} \pm 0.17$	-1.76 ± 0.24	-1.89 ± 0.17
Distance	0.034	0.031	0.033	**0.155**

Table 1: Average topic coherence values for each algorithm using the UMass coherence measure.



(a) (b)

Figure 2: (a) Topic flow matrix for the MTFM. (b) Topic flow matrices for each latent community of the LCTFM. Both were trained on the `/r/reddit.com` corpus.

corpus, which mentions many modern events and changes. The median topic coherence score per model is reported, and then these coherences are averaged over 10 trained models per algorithm. The MTFM has the most coherent topics using this measure, while the Bigram Topic Model (BTM) has the worst. The LCTFM and Mixture of Unigrams nearly tie for coherence in this case. While the LCTFM has much better predictive capabilities, it is not as strong at providing human-interpretable topic transition trends. Conversely, the MTFM can produce better interpretable topics with a predictive accuracy at the level of the Mixture of Unigrams.

One of the limitations of a single-topic approach is that many topic distributions can end up being similar to each other. The average Euclidean distance between topics are shown in Table 1. In comparison to the Biterm Topic Model, the Mixture of Unigrams, MTFM and LCTFM had a much shorter distance between topic distributions.

4.4 Topic flow characteristics

Figure 2 compares the transition matrix for the MTFM to one of the transition matrices for a community in the LCTFM. There are two notable differences between the transition matrices: First, the MTFM is more likely to have high self-transition probabilities, while the transition matrices in the LCTFM had only high self-transition probabilities in a small number of cases. Self-transition probabilities in both models were generally higher. This is understandable, since if people are discussing some topic, they are likely to Second, the transition matrices in the LCTFM were likely to have one or two topics that they commonly transitioned to; this is visualized as darker columns in the included figure. Many conversations were likely to revolve around a particular topic – for example, many discussions started with comments about some aspect of an article. Discussions would often diverge from that original topic, but we noticed that it would return after some time. Many other times, discussions would trend towards popular topics. In the `/r/politics` dataset, for example, many conversations would trend towards discussions about the results of the presidential race or ongoing popular issues. Such topic trends represent a significant but unaccountable factor to account for in the general MTFM. On the other hand, by categorizing each conversation into a group of documents, the LCTFM is capable of detecting contexts where conversations are likely to gravitate towards certain topics. This may be effective in the cyberbullying domain, where early conversation trends can be a strong indicator of escalation into harassment or other detrimental behavior.

5 Conclusion

This work presents a new model of topic flow with explicit document grouping information, the Latent Community Topic Flow Model (LCTFM). The LCTFM assumes that groups of documents have their own topic flow characteristics – where transitions may be determined by the document's author or the context in which a conversation takes place. Documents with similar conversational topic trends are collected into latent communities, which are used to enhance topic detection and flow. When compared to other models such as the Biterm Topic Model, the Mixture of Unigrams, and the conversation-aware Markov Topic Flow Model, LCTFM significantly outperforms all for predicting future conversation trends. One use-case for this work can be in the detection of online harassment or toxic comments, where conversations can escalate to undesirable outcomes.

One rigorous method for evaluating an unsupervised clustering algorithm is to test its effectiveness on a supervised classification problem. While there are a large number of labeled corpora available with a rich amount of metadata, there are only a handful of corpora with labeled social media conversations that preserve the conversation ordering. We plan to manually collect and label a social media dataset to further validate the LCTFM. An issue with our models was the high similarity between topic distributions compared to other approaches. The single-topic approach suffers when certain words become common across all topics (one example for our model was the word "people"). Future work will examine methods for filtering out such common words during model fitting.

References

[Barlier *et al.*, 2016] Merwan Barlier, Romain Laroche, and Olivier Pietquin. A stochastic model for computer-aided human-human dialogue. In *Interspeech 2016*, volume 2016, pages 2051–2055, 2016.

[Bickmore and Cassell, 2001] Timothy Bickmore and Justine Cassell. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 396–403. ACM, 2001.

[Blei and Lafferty, 2006] David M Blei and John D Lafferty. Dynamic Topic Models. *International Conference on Machine Learning*, pages 113–120, 2006.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[Blei *et al.*, 2017] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 2017.

[Daumé, 2009] Hal Daumé. Markov random topic fields. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, (August):293–296, 2009.

[Gutnik and Kaminka, 2004] Gery Gutnik and Gal Kaminka. Towards a formal approach to overhearing: Algorithms for conversation identification. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 78–85. IEEE Computer Society, 2004.

[Hofmann, 1999] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57. ACM, 1999.

[Liu *et al.*, 2009] Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. Topic-link LDA: joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 665–672. ACM, 2009.

[Mimno *et al.*, 2011] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 262–272, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[Nallapati *et al.*, 2011] Ramesh Nallapati, Daniel McFarland, and Christopher Manning. TopicFlow model: Unsupervised learning of topic-specific influences of hyperlinked documents. *Journal of Machine Learning Research*, 15:543–551, 2011.

[Neapolitan and Others, 2004] Richard E Neapolitan and Others. *Learning Bayesian Networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[Ritter *et al.*, 2010a] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations. *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (June):172–180, 2010.

[Ritter *et al.*, 2010b] Alan Ritter, Oren Etzioni, and Others. A Latent Dirichlet Allocation Method for Selectional Preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434. Association for Computational Linguistics, 2010.

[Röder *et al.*, 2015] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eight International Conference on Web Search and Data Mining, Shanghai, February 2-6*, 2015.

[Rosen-Zvi *et al.*, 2010] Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, 28(1):1–38, 2010.

[Steyvers *et al.*, 2004] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, (1990):306, 2004.

[Sun and Gao, 2008] Congkai Sun and Bin Gao. HTM : A Topic Model for Hypertexts. *Empirical Methods in Natural Language Processing*, (October):514–522, 2008.

[Toutanova *et al.*, 2003] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

[Wallach *et al.*, 2009] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112. ACM, 2009.

[Wang *et al.*, 2012a] Chong Wang, David Blei, and David Heckerman. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*, 2012.

[Wang *et al.*, 2012b] Yu Wang, Eugene Agichtein, and Michele Benzi. TM-LDA: Efficient Online Modeling of Latent Topic Transitions in Social Media. *Kdd*, page 123, 2012.