User Level Multi-Feed Weighted Topic Embeddings for Studying Network Interaction in Twitter

Pujan Paudel, Amartya Hatua, Trung T. Nguyen and Andrew H. Sung

{pujan.paudel, amartya.hatua, trung.nguyen, andrew.sung}@usm.edu
School of Computing Sciences and Computer Engineering
The University of Southern Mississippi
Hattiesburg, MS 39406, U. S. A

Abstract. Over half a billion tweets on a wide range of topics are posted daily by hundreds of millions of Twitter users. Insights of user behavior and network interactions can be applied to practical applications like targeted advertising, viral marketing, political campaigns, etc. In this paper, we propose a Multi-Feed Weighted Topic Embeddings (MFWTE) model to study user network interaction and topic diffusion patterns on Twitter. Our method extracts topic embeddings from multiple views of a Twitter user feed and weights them according to their content authoring roles, where the authored tweets, replied tweets, retweeted tweets, and favorited tweets are the views we separate for constructing the embeddings. We test the proposed method using two different topic modeling algorithms: i) Latent Dirichlet Allocation ii) Twitter-Latent Dirichlet Allocation. The users in our study are divided into multiple hierarchies based on their activity composition regarding individual topics, and the effectiveness of MFWTE is evaluated in the multi-hierarchical setting. The performance of our method on friendship recommendation and retweet behavior prediction task is evaluated using various ranked retrieval measures. The results indicate that our MFWTE method for topic modeling of Twitter users improves over various previous baselines. We conclude our work by applying the proposed model, MFWTE to discover various information diffusion patterns on Twitter.

1 Introduction

Microblogging platforms, such as Twitter and Reddit, have emerged as the primary platform on the internet for users all around the world to engage in discussions over a large variety of topics. Users can use Twitter to associate any tweet with any number of hashtags which allows the platform to aggregate large volumes of related tweets in real time. Hashtags in Twitter serve as an important explicit clustering mechanism that helps to recommend users and topics of similar interest in the platform. However, the use of hashtags as a primary mechanism for topic detection or any other interaction study would be

inconsistent and highly effective across Twitter users: In a previous study, Boyd et al. [9] have shown that only 5% of tweets contain a hashtag. Also, a group of users might interact about a topic using a set of different hashtags. Studying link formation and information diffusion pattern of users and their tweets within the wider Twitter network using topic models open a wide area of applications for targeting advertisements, user recommendations, and network analysis.

The most popular topic model is Latent Dirichlet Allocation (LDA) [6] which has grown quite popular for modeling large document collections with large text lengths. Traditional LDA models fail when applied over short and imbalanced texts with relatively shorter document lengths and skewed topic-word distribution. To limit the discrepancy of LDA models over short texts, various solutions have been proposed. Weng et al. [1] and Hong et al. [2] tried aggregating all the tweets of a user as a single document to account for short document length. Zuo et al. [3] performed modeling of distributions over topics using word co-occurrence matrix to alleviate the problem for short texts. Zhao et al. [4] proposed a model, Twitter-LDA by assuming the association of a single tweet with a single topic.

Embedding vectors have been used successfully in Natural Language Processing tasks such as sentiment detection in Tang et al. [7], and Text Classification and Neural Translation in Zou et al. [8]. Compared to generating embedding vector for texts, generating embeddings for multi-view data is non-trivial as it requires dealing with different modalities and distributional properties of the views. Benton et al. [5] proposed a weighted variant of Generalized Canonical Correlation Analysis to learn multiview embeddings of Twitter users. We were motivated by the effectiveness of their method in learning embeddings for the tasks of user engagement prediction, friendship prediction, and demographic attribute prediction.

User recommendation problem in Twitter space has been investigated before using various approaches. Graph-based approaches such as in Armentano et al. [12] consider topological position of users in network graph (followers as well as friends) to recommend potential followers. In the same work, they also propose content-based recommendation by comparing tweet content of the users social graph. Garcia et al. [12] propose weighted content-based recommendation method by identifying popularity, activity, location, mutual friends and tweet contents as features. Golder et al. [13] investigates structural approaches for user recommendation by using features like reciprocity, shared interests, shared audience and filtered people. Twittomender, a recommender system proposed by Hannon et al. [14] used an ensemble of profiling strategies, both content-based, and collaborative-filtering based, to recommend user profiles to follow. Experiments have shown that a combination of collaborative methods is more precise than individual content-based methods [15]. Similarly, sentiment-based [17] approach has also been applied for Twitter user recommendation.

Pennacchiotti and Gurumurthy [18] investigated topic models for social media user recommendations using an adapted user level LDA model, replacing

documents with users' Twitter stream. Their LDA system significantly outperformed the TF-IDF representations of users' tweets, demonstrating the applicability of topic models for capturing user-level behavior. Similarly, Ramage et al, in [19] used Labeled-LDA [20] to characterize Twitter users using topic-models. They demonstrated the weighted combination of TF-IDF model of user tweets and Labeled-LDA together performing well in the task of user-recommendation.

Most of the work done on retweet link prediction use similarity matching between the source profile and target profile of a user, or matching score between re-tweetable tweets under study and users topic of Interest. Multiple approaches have been adopted to create the user profiles of a user for similarity matching. Xu and Yang [22] proposed TF-IDF based bag of word profiles for each user on their similarity based model for retweet prediction. The work by [23] studied the information sharing strategies of users in online social networks under the strategical features of Interest Matching, Linguistics, Information Trustability and Information Freshness. The evaluation of strength of features in the same study reported the TF-IDF weighted Bag Of Words (BOW) similarity of reposted tweets performing as the most powerful signal on predicting user retweet behavior. They also compared the optimal information strategy model with LDA topic models and reported LDA models outperforming TF-IDF models using retweet strategy as features.

Most of the work related to topic modeling in the Twitter environment has focused on inferring topic distributions of individual isolated tweets, deviating from user topic profiling view-point. While [18] and [19] are the closest to our work, the scale of their evaluation sizes is relatively small, as they use sample size 8 and 10 respectively for positive/negative test users for evaluating experimental results. The evaluation methods used by our study use Information Retrieval metrics like Precision @K, Recall @K, Mean Reciprocal Rank (MRR) in comparison to traditional machine learning metrics, like accuracy and ROC curve used in previous works. The behavior of topic models and their effectiveness on network analysis on a much larger evaluation size needs to be explored more for its possible application in Big Data systems. The objectives of this work are as follows: 1) Comparing the performance of different topic modeling algorithms on user-level distributions for the task of network analysis, 2) Studying the effects of dividing Twitter feed into multiple views, based upon their content authoring source, 3) Evaluating the advantages of using weighted embeddings over non-weighted topic embeddings, 4) Dividing the study users into multiple hierarchies based on their topical activity and studying the effect of weighted topic embeddings on them, 5) Discovering information Diffusion patterns in user networks through the Multi-Feed Weighted topic models.

The paper is organized as follows: Section 2 discusses the data collected and methods applied in our work; Section 3 covers the experimental setup for our research; Section 4 highlights our results on the weighted multi-feed topic embeddings for friendship recommendation and retweet link prediction; Section

Table 1: Topics and Hashtags Used

Topic	Hashtags
Animal Rights	#animalrights, #animalabuse, #huntingkills, #saveanimals
Domestic Violence	#metoo, #domesticviolence, #sexualviolence, #violenceagainstwomen
Book Lovers	#bookworm, #books, #booklover, #bibilophile, #amreading
Net Neutrality	#netneutrality, #savetheinternet
Mental Health	#anxiety, #mentalIllness, #depression, #mentalhealth, #suicideawareness

5 investigates the findings and discusses contribution of the research for practical applications; Section 6 concludes our work.

2 METHODOLOGY

This section describes our methodology. We begin with the description of the dataset that we collected, the pre-processing pipeline, followed by the different views and embeddings of Twitter feed that we formulated and the topic modeling strategies that were applied.

2.1 Description of Dataset

Twitter lacks an explicit concept of topic space on their system. We compiled an individual topic as a collection of related hashtags. For our study, we selected five random topics and extracted five most common hashtags related to these topics, by observing the tweets associated with the users who tweeted multiple-tweets around those topics. The topics and their related hashtags used in our study are presented in Table 1.

We downloaded 2000 users each who tweeted at least more than 10 tweets under any single topic described in Table 1. While downloading the users, users with verified accounts and users with non-English profiles were removed. For each user, we collected 400 of their most recent tweets. We applied various pre-processing steps on these tweets to improve the quality of topics learnt by the models. We removed emoticons and other special characters which are a common source of noise on Twitter data. We removed all tweets not authored in English language, low-frequency words, stop words, HTML tags, and URLs from the tweets. We converted our entire vocabulary to lower cases, lemmatized them and expanded common English contraction words. Most importantly, we removed all

the hashtags from the tweet texts, to ensure that the hashtags wont influence the learned topic distributions.

For network analysis of users friendship and retweet graph using the learned topic embeddings, we expanded social links of users in our study by downloading user information of 1500 of their followers and 1500 of their friends. Again, we collected 400 of their most recent tweets and passed them through the same pre-processing pipeline as described above.

2.2 Topic Modeling

We applied two different topic modeling algorithms on the tweets of users passed through pre-processing pipeline: traditional LDA model and Twitter-LDA model. From here on, we refer to traditional LDA model as Canonical LDA. We used the implementation of [6] for Canonical LDA while [4] for Twitter-LDA. We subjected both of the topic models under identical parameters of Dirichlet prior for Document-Topic distribution(α), and Dirichlet prior for Topic-Word Distribution (β), of 0.5 and 0.01 respectively. The number of topics was set to 6, one more than the number of actual topical classes, to account for the background class inferred by Twitter-LDA. For both of the cases, Gibbs sampling was applied for model parameter estimation. We planned on using WNTM [3], but due to extreme memory consumed by the word occurrence matrix of WNTM, we were unable to apply this model to our study.

2.3 Multi-Feed Topic Modeling

To capture powerful representations of Topic Embeddings, we broke down Twitter feeds of individual users into multiple views based on their content authored sources: A) **Authored View:** View composed with the tweets primarily authored by the users. B) **Replied View:** View composed with tweets sent as a reply to other tweets. C) **Retweeted View:** View composed with tweets forwarded by the user. D) **Favs View:** Views composed with tweets favorited (liked) by the user. We are aware of community detection techniques using such multi-view approaches of data for community identification before. Greene and Cunningham [21] construct a heterogeneous collection of content-based views, incorporating views like tweet content, list text, mentions, retweets to produce unified graph representations for the task of community detection. Kwak et al. [10] compare trend analysis of users on their large-scale work of Twitter by observing the behavior of trending topics in 'Singleton', 'Reply', 'Mention' and 'Retweet' views. A similar methodology of differentiation between content source of tweets was done in the work of [16] where the comparison model selected profiling strategy of a user as either an 'author' or a 'retweeter' using topic similarity scores based on past tweets and retweets. Our work differs from the work of [16] in that the authors separated tweet source as a noise removal strategy for behavior prediction, while we are incorporating multiple views to capture dynamic content creation and sharing

mechanisms of a Twitter user, extending beyond hard profile limitations of an 'author' or a 'retweeter' and allowing multiple content diffusion roles to be studied. We ran the same set of topic detection algorithms with identical parameters as discussed in Section 2.2 on the Multi-Feed topic models. For rest of the work, we refer the topic models of unseparated, traditional twitter feed as **single-feed** embeddings while the separated feed as discussed before as **multi-feed** embeddings.

2.4 Multi-Feed Weighted Topic Modeling

We felt the need to weigh the multi-feed topic embeddings of our user feeds discussed in Section 2.3 with different weightage in order to produce efficient retrival results. We used the implementation by [5] for generating weighted embeddings of our multi-feed topic models, to propose our final model, MFWTE. We explored multiple view lengths {15, 20, 40, 100}, multiple view weightages {2, 5, 10, 20, 40, 80} and performed grid search over the model to investigate the effect of multi-view weighted topic embeddings.

2.5 Hierarchical Study of Twitter users

We divided the users in our study into three hierarchies based on a user's Twitter activity towards topical distributions. We used results of the same topic modeling algorithm used in Section 2.2 to extract the hierarchies of users. The hierarchies of users were defined as follows: A) **Tier 1 Users:** are the primary content creator for a topic, who tweet 85-95% about an individual topic. B) **Tier 2 Users:** are secondary content creators for a topic, who compose relatively lesser tweets about a topic while their topic composition being 70-85% on an individual topic. C) **Tier 3 Users:** very rarely author tweets by themselves related to a topic. Their Twitter activities amount to 50-60% about an individual topic. The users who fall under this tier generally have a multitude of interests and tweet almost equally about multiple topics. A similar distribution of study users into multiple buckets can be observed in [18], where they divided their experimental setup into head, torso and tail for investigating topic models for social media recommendation.

Table 2: Distribution of user tiers and evaluation set size per user for friendship recommendation

Evaluation Tier	Friend Sample	Non-Friend Sample	Dev-set
1st Tier	530	530	1534
2nd Tier	344	344	1120
3rd Tier	419	419	1463

Table 3: Distribution of user tiers and evaluation set size per user for retweet behavior

Evaluation Tier	Retweet Sample	Non-retweet Sample	Dev-set
1st Tier	123	250	880
2nd Tier	106	220	1090
3rd Tier	150	289	1200

The formulation of our tiers differs from their division of buckets as their method divided users according to the number of followers, while our division is based on the topical activity of the users. The motive behind dividing the user base into multiple tiers was to study the effect of topic modeling on network prediction tasks separately on these multiple levels, as we observed the behavior of the topic models, and multiple views of it, vary amongst users who have different content authoring behavior over the platform.

3 EXPERIMENTAL SETUP

This section explains the framework of analysis that we used to evaluate our method. We selected the task of friendship recommendation and retweet link prediction to examine the performance of topic models on users network interaction. Following relationship and retweeting relationship have been proven to be closely correlated with users interest as reported by Weng et al. [1].

For evaluating our models on the task of friendship recommendation, we made our evaluation sets equally balanced on all our tests by including an equal count of friend (positive) samples and non-friend (negative) samples. Every user has a set of friend and non-friend users. Non-friend users are defined as the users who are followed by ten of another user's friends, but who are not followed by the user. Compared to distinguishing between friend/non-friends of a user, identifying non-retweet links between Twitter users is non trivial. If user A is following user B, and user A has retweeted more than 10 tweets from B, then we sampled a directed link from A to B under positive class. If user A is friends with user B, but has less than 10 retweet links coming from user B and more than 10 friends who have 10 or more retweet links from user B, user B is sampled under the negative set. For retweet behavior prediction tasks, we followed prior adopted approaches by multiple works, [23] and [**?**], to sample nearly double non-retweet (negative) samples compared to retweet (positive) samples. For both of the link prediction tasks,we isolated the Dev-set used to learn the multi-view embeddings from those that were used in the evaluation sets, to remove the possibilities of biases on the learnt embeddings For every tier of users, we held out random 100 target users; for each target user, we selected a positive set of users and a negative set of users. The distribution of evaluation set size for individual user, tier-wise, is displayed in Table 2 and Table 3 for the two link prediction tasks, respectively. As seen in the tables, our evaluation

size is much larger compared to those of previous studies. Scoring was done by ranking the compared cosine distance between the topic vectors of target users and users of the evaluation set. Evaluation was done using ranked retrieval metrics of Precision @K, Recall @K, as well as Mean Reciprocal Rank (MRR). A friendship connection between Twitter users is scored as a positive "hit" for the task of friendship recommendation while a directed retweet link between two users is scored as a positive "hit" for the task of retweet link prediction. The size of our evaluation sets for different studies as well as for different tiers of users was different, so we modified the metrics of Precision @K and Recall @K slightly to account for the percentage of positive class size present in the evaluation set. For example, Recall @0.5 refers to Recall @50% of the count of positive samples present in the evaluation set.

4 RESULTS

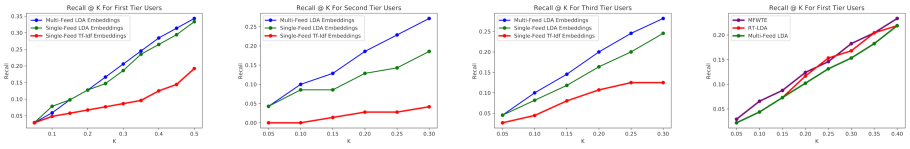
This section discusses the results of our experimental study. We begin with the comparison of single-feed and multi-feed topic embeddings followed by the performance of different topic detection algorithms. We report the performance of weighted multi-feed topic embeddings. This section is concluded by the results on the application of weighted multi-feed embeddings to discover various topic diffusion patterns.

4.1 Comparison of Multi-Feed with Single Feed Twitter Topic Embeddings

Before comparing the performance of multi-feed and single-feed topic model using ranked retrieval methods, we did an initial inspection of the quality of topic vectors using machine learning classifiers and machine learning evaluation metrics. For our proposed testing framework of friendship recommendation as discussed in Section 3, we trained a Support Vector Machine(SVM) using linear kernels for multi-feed and single-feed topic embeddings. We used the results from the topic modeling algorithms discussed in Section 3 as the input feature vectors. Evaluation was done using 10-fold Leave One Out cross-validation. For both of the link prediction tasks of friendship recommendation and retweet prediction, the multi-feed topic embeddings reports higher values of F1 score and accuracy over the baseline single-feed embeddings in all of the three hierarchies of users. The comparison tables of F1 score and accuracy is not displayed for the sake of brevity.

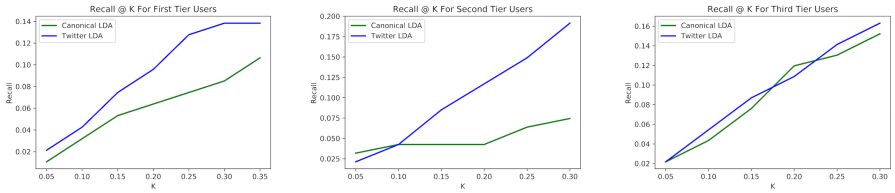
Next, we evaluated our Multi-Feed Embeddings using Ranked Information Retrieval measures. For the experimental framework of friendship recommendation, we compared Multi-Feed Topic Embeddings with the baseline of Single-Feed Topic Embeddings and TF-IDF embeddings of user tweets. Fig. 1 shows the recall @K curve as a function of number of recommendations. Multi-Feed approach outperforms the Single-Feed approach as well as the TF-IDF Embeddings while evaluating under Precision @K, Recall @K and

Mean Reciprocal Rank (MRR). This holds true for all the tiers of users we had divided, so we learned that dividing the tweets in Multi-Feed views, based upon their activity source, helps in improving the topic models for user behavior. We repeated identical Ranked Retrieval evaluation for retweet link prediction using topic models by comparing Multi-Feed Topic Embedding against multiple baselines of Single-Feed Topic Embedding, Topic Embedding of retweets collection of a user, and Bag Of Words (BOW) Weighted TF-IDF embeddings of a users tweet. The topic embeddings performed better than TF-IDF embeddings as expected. But, as seen in Figure 1, neither the Multi-Feed Embeddings or Retweets topic Embeddings had a clear advantage in performance over the other as the value of Recall @K fluctuated with varying value of K.



(a) (b) (c) (d)

Fig. 1: Recall @K for friendship recommendation in a) First Tier user, b) Second Tier user, c) Third Tier user, d) Recall @K for retweet prediction



(a) (b) (c)

Fig. 2: Recall @K for canonical LDA and Twitter LDA a) First Tier user b) Second Tier user c) Third Tier user

We investigated two different types of topic models: Canonical LDA and Twitter LDA. Our analysis of the two modeling algorithms on the task of friend-ship recommendation shows that Twitter LDA outperforms the Canonical LDA model under Recall @K metric for all the tiers of our user division. Figure 2 shows the Recall @K curve for the two topic models as a function of number of recommendations for all three tiers of users. There were identical results reported in the task of Retweet behavior prediction. The figures are left out for sake of brevity. In [18], the experimental results

concluded that the user-level topic models are effective over tweet-level topic models. Though an indirect comparison under different dataset and different user-level tweet model (Twitter-LDA compared to their Labeled-LDA), our results disagree with their findings.

4.2 Comparison of Multi-Feed Weighted Topic Embeddings with Non-weighted Embeddings

After verifying our initial proposal for the topic embeddings extracted from multiple content authoring sources in Section 4.1, our next question was if subjecting those learnt embeddings to weighted model would provide even more performance boost. Weighing models are important because of the different content authoring activity exhibited on different tiers of users and across different views of user feeds we had formulated. We did grid search over multiple weights and their combinations over multiple views to find the optimal weights for different tiers of users. Another example where Multi-View Embeddings have been used before was by Benton et al. [5] where they used multiple views (Ego Tweet, Friend tweets, Followers tweet) for the task of friendship prediction. For our evaluation, we used the best performing Multi-Feed topic embeddings from Section 4.1 and TF-IDF weighted Bag Of Words (BOW) representation of the multiple-content sources tweets (Authored tweets, Replied tweets, Retweeted tweets and Favorited tweets) as discussed above as baselines. We performed identical grid searches of weights for the TD-IDF embedding baseline used for comparison. Comparison of Recall @K and Precision @K performance as function of number of recommendations is given in Figure 3. We observed that the Multi-Feed Weighted Topic Embeddings (MFWTE) outperforms the best performing model from Section 4.1 and Multi-Feed TF-IDF Embeddings under Recall @K Metric. This verifies our proposed idea of Multi-Feed Weighted Topic Embeddings having an advantage over non-weighted embeddings.

We evaluated MFWTE for the task of retweet link prediction using identical experimental settings and similar baselines of best Multi-Feed Topic Embeddings from Section 4.1. For this study,we added Topic Embeddings of Retweeted Tweets and Multi-Feed TF-IDF embeddings as our baselines. It can be observed from the comparative analysis of the Precision @K and Recall @K curves in Figure 4 that the MFWTE outperforms all other baselines for retweet link Prediction in the first and second tier of users. For the third tier of users, the Retweets Topic Embeddings perform the best. This can be explained by the topic composition instability of third tier users, who are generally passive content retweeters of multiple topics in Twitter and their retweet behavior goes uncaptured even by the powerful Multi-Feed Topic Embeddings.

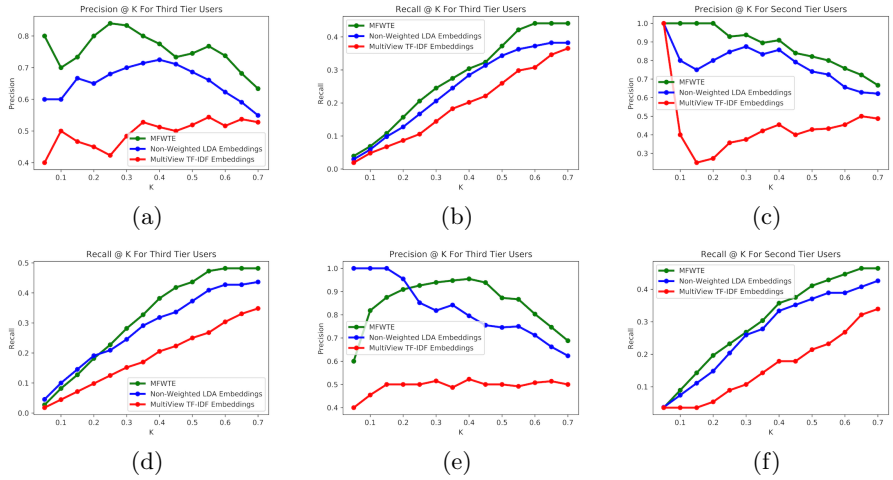


Fig. 3: Friendship recommendation comparison of MFWTE with baselines a & b (First Tier), c & d (Second Tier), e & f (Third Tier)

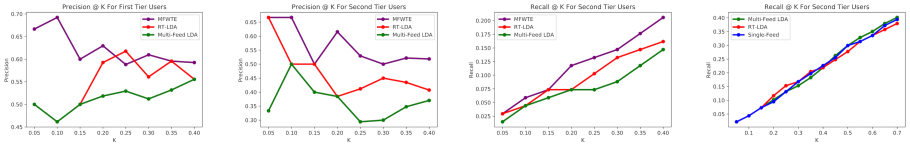


Fig. 4: Retweet prediction comparison of MFWTE with baselines a & b (First Tier), c & d (Second Tier)

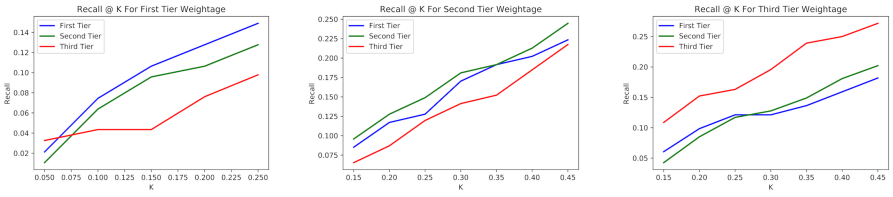
4.3 Identifying Topical Information Diffusion Patterns Using MFWTE

In Section 4.2, we showed that weighing the embeddings improved the efficiency of our learnt topic models. We were motivated to find topical information diffusion patterns in the different hierarchy of users using the weighted embeddings. We applied MFWTE in investigating parameter combination space of topic embeddings at different content authoring behaviors throughout the different tiers of users. With α being the weight for **Authored View**, β being the weight for **Replied View**, γ being the weight for **Retweeted View**, and θ being the weight for **Favs View**, the weight combinations we used for the study of three different topical diffusion patterns is depicted in Table 4.

For Tier 1 of users, we proposed the primary content creators would engage in replying to topical tweets at a much higher rate than the other two tiers of users. Weightage combination C1 highlights our diffusion pattern for this case

Table 4: Weightage Parameters of Different Views on Different Tiers of Users

Weightage Combination	α	β	γ	θ
C1	1	40	1	1
C2	20	20	1	1
C3	1	1	20	20



(a) (b) (c)

Fig. 5: Recall values of MFWTE under a) C1 b) C2 c) C3

study. We subjected the test set of the other two tiers under same weight combination, which showed that C1 and the performance on number of positive friendship recommendation drops with the decrease in tier level. This validates our formulation of weight in Replied Views more for the first tiers of users. This is explained by the observation that non-topical users do not engage much in replying tweets related to a topic. We repeated similar experiment on Tier 2 of users with weight combination C2, proposing they engage in authoring the Tweet and Replying Tweets almost equally, while still being active in terms of authoring contents related to a topic compared to forwarding them. Weightage combination C2 highlights our diffusion pattern for this case study. We evaluated the performance of other tiers on the similar weight C2 and noted that the highest recall for positive friendship recommendation is observed for second tier users the most, followed by the first-tier users, which confirmed our initial assumption of weight attribution towards the second-tier users. Similarly, Tier 3 were content distributors; users who created less content for the topics themselves, but retweeted and favorited topical tweets. Weightage combination C3 highlights our diffusion pattern for this case study. When we subjected the three tiers of users under the weight combination of C3, it was observed that positive friendship recommendation performance in the third tier of users perform exceptionally well while the performance drops for the other two tiers with increase in tier level. The Recall @K curve of different information diffusion patterns and the effectiveness of MFWTE in capturing them is shown by Figure 5. Thus, with the application of MFWTE using case specific weighted embeddings, we were able to demonstrate different content topic diffusion patterns for the different tiers.

5 DISCUSSION

Users in the Twitter platform demonstrate highly complex activity of interaction, thus their authoring, replying and forwarding behavior in tweets are highly significant to determine their topics of interests and possible friendship connections. We divided a single stream feed of users into multi-view feeds, based upon their content authoring sources and were able to build better topic models for predicting friendship links and modeling retweet behavior in Twitter. The idea of segregating tweets with this configuration allows us to capture information diffusion in a highly dynamic environment like Twitter and create efficient collaborative filtering methods for user recommendations and information propagation. The four views which we have formulated can be extended to any number of activity sources, also extending to multiple social identities of an individual user (like Facebook) to build quality topic models. We applied two different variants of topic modeling algorithm on our datasets and discovered that Twitter-LDA has higher performance in the task of friendship recommendation than the LDA model designed for traditional documents. Thus, Twitter-LDA can be pursued as more efficient topic-modeling algorithm in other Big-Data analysis tasks for Twitter.

All of our Multi-Feed Topic Embeddings (both weighted as well as unweighted) perform better than their TF-IDF baselines. This result agrees with the findings by [18] where their adapted LDA system outperformed the TF-IDF baseline with statistical significance. Our results reinforce the claim that topic models are indeed good representations of user-level interests by demonstrating their efficiency in two link prediction tasks. The efficient performance of our final proposed model under Big-Data scale Information-Retrieval evaluation metrics (Ranked Retrieval when compared to ROC Curve used in [18]) confirms the application of Multi-Feed Weighted topic models as good representation of user level interests.

Improving over the Multi-Feed topic embeddings, our final model, MFWTE demonstrates even better results in all tiers of the user base we had formulated, opening up a wide and interesting area of application in high-impact marketing campaigns. MFWTE allowed learning highly dynamic topic embeddings based upon the tiers of user we are interested in targeting, as well as to capture different activity variance of Twitter users over multiple modes of interaction. Learning dynamic weights will help for improving targeted information outreach among different types of user bases. The topical information diffusion patterns we studied using MFWTE can be extended to analyze the spreading behavior of viral topics over different forms of interaction in a platform, as well as across different tiers of users in the platform. One such possible use case of it is the weighted embeddings of "Favs" view that we have learnt from our models. They can help identify the topic interest of users who may not be an enthusiast on a topic in terms of authoring them, or forwarding them but who are latent observers of the activities related to that topic. This type of users, who are quite common in the platform, might be reached for marketing and information campaigns, which otherwise might have remained unnoticed.

6 CONCLUSIONS

The main contribution of our paper is a weighted, multi-feed topic embedding which better captures topical interests of a users tweets and demonstrates better performance in discovering friendship connection and modeling retweet behavior on a large scale Twitter user network than previous models.

Our proposed methodology of segregating the user feed into multiple views based upon their content authoring sources helped us to capture the dynamics of the complex Twitter system and build better topic embeddings than traditional Twitter topic models. Further validation of the effectiveness of our model was done by evaluating them using different topic modeling algorithms and under different tiers of users. Being motivated by the effectiveness of multi-feed embeddings, we learned weighting parameters of the embeddings by grid-search over the parameter combination space improved over the non-weighted topic models. Our proposed final model Multi-Feed Weighted Topic Embeddings performs the best in Ranked Retrieval experiments, taking advantage of multi-view feeds as well as weighted embeddings learnt from WGCCA at the same time. Finally, by applying the MFWTE model on different tiers of our user-sets, we discovered multiple topic-level content authoring patterns of the users.

For future work, the multi-feed weighted models can be tested on other variants of topic modeling algorithms, like [3] and [19] requiring an extensive amount of working memory. Examining the effectiveness of these learnt topic models to community detection and other network analysis problems is another vital direction for future work.

References

[1] Weng, J., Lim, E.P., Jiang, J. and He, Q., 2010, February. Twitterrank: finding topic-sensitive influential twitterers. In Proceedings of the third ACM international conference on Web search and data mining (pp. 261-270). ACM.

[2] Hong, L. and Davison, B.D., 2010, July. Empirical study of topic modeling in twitter. In Proceedings of the first workshop on social media analytics (pp. 80-88). ACM.

[3] Zuo, Y., Zhao, J. and Xu, K., 2016. Word network topic model: a simple but general solution for short and imbalanced texts. Knowledge and Information Systems, 48(2), pp.379-398.

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

[5] Benton, A., Arora, R. and Dredze, M., 2016. Learning multiview embeddings of twitter users. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Vol. 2, pp. 14-19).

[6] Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.

[7] Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T. and Qin, B., 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In Proceedings

of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 1555-1565).
[8] Zou, W.Y., Socher, R., Cer, D. and Manning, C.D., 2013. Bilingual word embeddings for phrase-based machine translation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1393-1398).
[9] Boyd, D., Golder, S. and Lotan, G., 2010, January. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In System sciences (hicss), 2010 43rd hawaii international conference on (pp. 1-10). IEEE.
[10] Kwak, H., Lee, C., Park, H. and Moon, S., 2010, April. What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World wide web (pp. 591-600). AcM.
[11] Armentano, M.G., Godoy, D. and Amandi, A.A., 2011, July. Recommending information sources to information seekers in twitter. In International workshop on social web mining.
[12] Garcia-Gavilanes, R.O.G.G. and Amatriain, X., 2010. Weighted content based methods for recommending connections in online social networks.
[13] Golder, S.A., Yardi, S., Marwick, A. and Boyd, D., 2009, July. A structural approach to contact recommendations in online social networks. In Workshop on search in social media, SSM.
[14] Hannon, J., McCarthy, K. and Smyth, B., 2011, April. Finding useful users on twitter: twittomender the followee recommender. In European Conference on Information Retrieval (pp. 784-787). Springer, Berlin, Heidelberg.
[15] Kywe, S.M., Lim, E.P. and Zhu, F., 2012, December. A survey of recommender systems in twitter. In International Conference on Social Informatics (pp. 420-433). Springer, Berlin, Heidelberg.
[16] Syeda Nadia Firdaus, Chen Ding, and Alireza Sadeghian. 2016. Retweet prediction considering user's difference as an author and retweeter. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '16). IEEE Press, Piscataway, NJ, USA, 852-859.
[17] Gurini, D.F., Gasparetti, F., Micarelli, A. and Sansonetti, G., 2013. A Sentiment-Based Approach to Twitter User Recommendation. RSWeb@ RecSys, 1066.
[18] Pennacchiotti, M. and Gurumurthy, S., 2011, March. Investigating topic models for social media user recommendation. In Proceedings of the 20th international conference companion on World wide web (pp. 101-102). ACM.
[19] Ramage, D., Dumais, S.T. and Liebling, D.J., 2010. Characterizing microblogs with topic models. ICWSM, 10(1), p.16.
[20] Ramage, D., Hall, D., Nallapati, R. and Manning, C.D., 2009, August. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1 (pp. 248-256). Association for Computational Linguistics.
[21] Greene, D. and Cunningham, P., 2013, May. Producing a unified graph representation from multiple social network views. In Proceedings of the 5th annual ACM web science conference (pp. 118-121). ACM.
[22] Xu, Z. and Yang, Q., 2012, August. Analyzing user retweet behavior on twitter. In Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on (pp. 46-50). IEEE.
[23] Nguyen, D.A., Tan, S., Ramanathan, R. and Yan, X., 2016, August. Analyzing information sharing strategies of users in online social networks. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 247-254). IEEE Press.