Twitter Topic Modeling for Breaking News Detection

Henning M. Wold[1], Linn Vikre[1], Jon Atle Gulla[1], Özlem Özgöbek[1,2] and Xiaomeng Su[3]

[1] *Department of Computer and Information Science, NTNU, Trondheim, Norway*

[2] *Department of Computer Engineering, Balikesir University, Balikesir, Turkey*

[3] *Department of Informatics and e-Learning, NTNU, Trondheim, Norway*

Keywords: Twitter, Topic Modeling, News Detection, Text Mining.

Abstract: Social media platforms like Twitter have become increasingly popular for the dissemination and discussion of current events. Twitter makes it possible for people to share stories that they find interesting with their followers, and write updates on what is happening around them. In this paper we attempt to use topic models of tweets in real time to identify breaking news. Two different methods, Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP) are tested with each *tweet* in the training corpus as a document by itself, as well as with all the *tweets* of a unique user regarded as one document. This second approach emulates Author-Topic modeling (AT-modeling). The evaluation of methods relies on manual scoring of the accuracy of the modeling by volunteered participants. The experiments indicate topic modeling on *tweets* in real-time is not suitable for detecting breaking news by itself, but may be useful in analyzing and describing news tweets.

1 INTRODUCTION

Social media networks facilitate communication between people across the world. Social networks not only make it easier for people to communicate, they also make it possible for the media to capture breaking news as they are emerging. Social media networks have been used to provide information in real-time about larger crisis situations such as earthquakes and tsunamis (Mendoza et al., 2010).

Twitter is one such social network and a microblogging service founded in 2006. As of 2014 the company reports having 284 million active users per month[1]. The service is focused on micro messages, called *tweets*, which are restricted to a length of 140 characters. In addition to posting *tweets* about anything, users can also follow other users.

In (Sakaki et al., 2010) it is examined how earthquakes could be detected using Twitter. In this research Twitter users are treated as sensors and the *tweets* as sensor data. By using this approach it was possible to detect 96% of the earthquakes with intensity of 3 or more occurring in the examined area. In (Hu et al., 2012) it is showed how the news of Osama Bin Laden's death spread on Twitter before the mass media could get the news confirmed.

These studies suggest that Twitter can be used effectively to detect breaking news before they are published in traditional news media. To do that *tweets* that can be considered news-worthy should be identified while disregarding the noisy ones. Here we use the term noisy as the *tweets* which are expressing personal matters. Subsequently, for the *tweets* detected in the first process, the *tweets* that are deemed untrustworthy need to be pruned. The challenges that we consider in this paper are:

- Finding a suitable topic modeling method for tweets,
- Continuous training of the model as we acquire new *tweets* and real-time processing of the issues above.

This paper concerns the issue of finding out newsworthy tweets and seek to find a possible strategy for collecting breaking news through Twitter using topic modeling techniques.

The proposed work in this paper is continuing as a part of the SmartMedia program[2] which was started in 2011 at Norwegian University of Science and Technology (NTNU) in close collaboration with Scandinavian media industry. With this program it is targeted to present online news in an effective and personalized way to the users (Gulla et al., 2014), as well as

[1] https://about.twitter.com/company

[2] http://research.idi.ntnu.no/SmartMedia

Wold, H., Vikre, L., Gulla, J., Özgöbek, Ö. and Su, X.
Twitter Topic Modeling for Breaking News Detection.
In *Proceedings of the 12th International Conference on Web Information Systems and Technologies (WEBIST 2016) - Volume 2*, pages 211-218
ISBN: 978-989-758-186-1

to build the context aware news exprериences based on deep understanding of text in continuous news streams (Ingvaldsen et al., 2015).

In Section 2 we present the previous research in the field of topic modeling in general, and on *tweets* in particular. Section 4 describes how we aim to find a suitable topic modeling technique to handle tweets in real time. After that, we evaluate the results of the different topic modeling techniques in Section 5. In Section 7, we discuss our results while Section 8 summarizes our findings along with the proposal of future work.

2 RELATED WORK

A topic model of a collection of documents is a trained statistical model that exposes abstract topics in the collection. Each document may concern multiple topics in different proportions, like a news article about pets that may write about topics like cats, dogs, and fish. The topics themselves are represented by high-frequency words that occur in the descriptions of the topics. Historically, topic modeling has been widely used to explore topical trends in large corpora that spans several years, analyzing for example political and social changes in important time periods.

In recent years, topic modeling has been increasingly utilized for analyzing corpora of *tweets*. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is one of the most widely used techniques for this analysis. There are several modifications and extensions proposed to LDA that improves its performance in social media settings in general, and for *tweets* in particular.

In (Rosen-Zvi et al., 2004) an extension to LDA called the Author-Topic Model (AT model) is proposed. In (Rosen-Zvi et al., 2010), it is showed that when the test documents contain only a small number of words, the proposed model outperforms LDA. This research was done on a collection of 1,700 NIPS conference papers and 160,000 CiteSeer abstracts. The work is done on abstracts which are shorter than normal documents but still longer than regular *tweets*.

(Hong and Davison, 2010) showed how training a topic model on aggregated messages results in higher quality learned models, yielding better results when modeling *tweets*. In this work all the messages of a particular user are aggregated before training the LDA model on them. This simple and straightforward extension to LDA gives a more accurate topic distribution than running LDA on each *tweet* individually.

In (Zhao et al., 2011) an empirical comparison is done between Twitter and the New York Times. Using an extension to LDA called Twitter-LDA is used, they concluded that Twitter could be a good source of news that has low coverage in other news media. The study also suggets that while Twitter users are not exceptionally interested in world news, they do help spread awareness of important world events.

Another method that has been used for topic modeling *tweets* is Hierarchical Dirichlet Processes (HDP) (Teh et al., 2006). HDP is a nonparametric Bayesian approach which is used to cluster related data and can be used to cluster *tweets* that have similar topics. (Wang et al., 2013) describes how HDP can be used to detect events occurring from *tweets* in real-time and shows how the clustering in HDP works on these *tweets*.

All of these studies suggest that there are several techniques and methods that perform well in different areas that can be utilized for our purpose of collecting breaking news from Twitter. Although there are studies that have experimented with similar ideas (Zhao et al., 2011), there are few who have tried to do this in real-time. Most studies have done this in semi real-time or looked at previous data to see if they could get an indication of whether or not it is possible to fetch potential breaking news using Twitter.

3 TWITTER NEWS DETECTION

When news-worthy events happen, people who witness it are often quick to post about the event to their social network feeds in general, and to their Twitter accounts in particular. In (Kwak et al., 2010) it was found that any retweeted *tweet* reached an average of 1,000 users. This means Twitter could be an interesting place for detecting breaking news as they are emerging. Unfortunately, many of the posts on Twitter are not news items, but rather address personal opinions and mundane status updates or similar, such as those found in figures 1 and 4. Other *tweets* are more serious and potentially related to news, such as figure 3. Lastly some *tweets* (e.g. figure 2) are entirely written in foreign languages often using non-latin alphabets. The first step in detecting breaking news is then to filter out the noise, leaving posts that are potential news for the analysis.

The main challenge of news detection on Twitter is the length restriction on *tweets*. A *tweet* can be a maximum of 140 characters long. Because *tweets* are

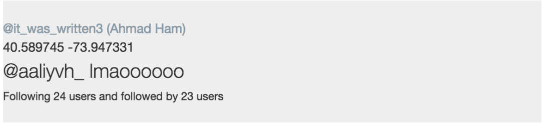
@it_was_written3 (Ahmad Ham)
40.589745 -73.947331
@aaliyvh_ lmaoooooo
Following 24 users and followed by 23 users

Figure 1: Example of short, non-serious, *tweet*.

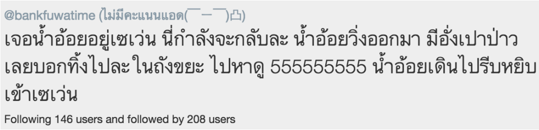
@bankfuwatime (ไม่มีคะแนนแอด(￣－￣)凸)
เจอน้ำอ้อยอยู่เซเว่น นี่กำลังจะกลับละ น้ำอ้อยวิ่งออกมา มีอั่งเปาป่าว เลยบอกทิ้งไปละในถังขยะ ไปหาดู 555555555 น้ำอ้อยเดินไปรีบหยิบเข้าเซเว่น
Following 146 users and followed by 208 users

Figure 2: Example of *tweet* written in a foreign language.

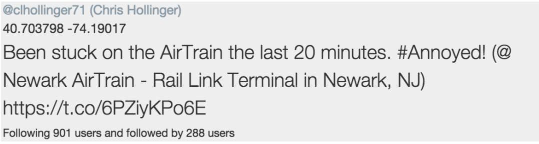
@clhollinger71 (Chris Hollinger)
40.703798 -74.19017
Been stuck on the AirTrain the last 20 minutes. #Annoyed! (@ Newark AirTrain - Rail Link Terminal in Newark, NJ) https://t.co/6PZiyKPo6E
Following 901 users and followed by 288 users

Figure 3: Example of long, serious, *tweet*.

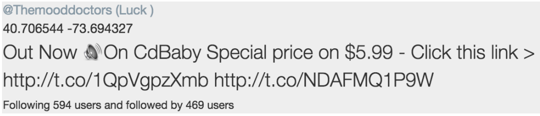
@Themooddoctors (Luck)
40.706544 -73.694327
Out Now 🔊On CdBaby Special price on $5.99 - Click this link > http://t.co/1QpVgpzXmb http://t.co/NDAFMQ1P9W
Following 594 users and followed by 469 users

Figure 4: Example of spam *tweet*.

so short and some posts might not directly use common keywords for describing an event, simply looking for certain keywords will result in only a handful of the actual news posts being detected. In addition, the list of keywords would have to be updated manually as the news context evolves over time.

Instead of using fixed keywords for detecting news on Twitter, we have focused on a few different topic modeling techniques enabling us to dynamically find potential news. We also combined the techniques with a news index. These techniques allow to update the models over time, ensuring that they stay relevant.

Topic modeling may enable us to identify a new tweet as news-related even though we do not have an exact list of keywords that constitute news. Moreover, the technique has the additional benefit of classifying tweets topic-wise and even associating them with some prominent topic words that may not actually appear in every tweet of this topic. The topic words provide useful summaries and the whole structure of topics associated with tweets produces a more flexible way of grouping tweets than traditional clustering.

4 METHODOLOGY

In this section, we introduce a set of methods for training topic models concerning Twitter. After clarifying the theoretical aspect of the LDA and HDP models we discuss how they can be used to topic model *tweets* in real-time.

4.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a machine learning technique that identifies latent information about topics in large collections of documents. LDA treats each document as a vector of word counts where each document is a probability distribution over some topics, and each topic is the probability distribution over a number of words. For each document in the collection the LDA algorithm picks a topic according to the multinomial distribution of words in the document. It then uses the topic to generate the word itself according to the topic's multinomial distribution and repeats these two steps for all the words in the document (Blei et al., 2003).

LDA can be defined formally as follows:

1. A number of topics t, documents d, and a length of a document N.
2. T distributions over the vocabulary where β_t is the distribution over words in topic t.
3. D distributions over topics where θ_d is the distribution of topics in document d.
4. The topic given for document d is z_d, and $z_{d,n}$ is the assignment of a topic to the nth word in document d.
5. The observed words in document d are given as w_d, where $w_{d,n}$ is the nth word in document d.

This gives us the formula (Blei, 2012) :

$$p(\beta,\theta,z,w) = \prod_{i=1}^{T} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) (\prod_{n=1}^{N} p((z_{d,n}|\theta_d)) p(w_{d,n}|z_{d,n,\beta}))$$

LDA has become one of the "state-of-the-art" topic modeling algorithms in the later years after it was presented in (Blei et al., 2003). The algorithm uses the "bag-of-words" principle to represent the documents, which is satisfactory when dealing with larger documents where it is possible to explore co-occurrences at the document level. With this it is possible to achieve a clear overview of all topics related to a document (Titov and McDonald, 2008), which in many cases is the ideal result.

The LDA model has been shown to work fairly well when topic-modeling *tweets*. Some studies, however, suggest that slight modifications to the LDA model, such as the AT model (Hong and Davison, 2010) and Twitter-LDA (Zhao et al., 2011) perform even better. One reason for this is the document length, which has a fairly large impact on the outcome of the topic modeling result.

Online LDA is a variation of LDA in which the model can be updated on the fly. As we are interested in topic modeling of a real-time stream of *tweets*, traditional LDA may not be sufficient for our purposes. For this reason, we have chosen to use *Online LDA*. So by using this model it is possible to update and continuously build upon the already existing model, which is based on the information that comes from the data stream. In our case that means new documents (that is *tweets*) will be added to the model continuously as they are received by the system.

4.2 Author-topic

Author-topic (AT) model is an extension of LDA. In this model, the content of each document and the interests of authors are simultaneously modeled. AT uses probabilistic "topics-to-author" model, which allows a mixture of different weights (θ and ϕ) for different topics to be determined by authors of a specific document (Rosen-Zvi et al., 2004). ϕ is generated similar to θ, where it is chosen from a symmetric Dirichlet(β).

The AT-model as defined above, would result in a very large memory footprint if directly implemented. The reason for this is that we would have to store all inbound *tweets* to be able to continuously update the documents written by each author, which would scale linearly in time. So instead of implementing it directly we replicate the approach of (Hong and Davison, 2010). This involves aggregating all the documents made by a single author into one document. In our case this means concatenating all the tweets of a unique user into one document.

4.3 Hierarchical Dirichlet Process

The Hierarchical Dirichlet Process (HDP) is a topic modeling technique for performing unsupervised analysis of grouped data. HDP provides a nonparametric topic model where documents are grouped by observed words, topics are distributions over several terms, and every document shows different distributions of topics. Its formal definition, due to (Teh et al., 2006) is:

$$G_0|\gamma, H \sim DP(\gamma, H)$$

$$G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0)$$

for each group j. Here G_j is the random probability measure of group j and its distribution is given by a Dirichlet process. This Dirichlet process depends on α_j, the concentration parameter associated with the group, and G_0. G_0 is the base distribution shared across all groups. This base distribution is given by a Dirichlet process as well, where γ is the base concentration parameter associated with the group and H is the base distribution which governs the a priori distribution over data items.

One limitation of standard HDP is that it has to crawl through all the existing data multiple times. Therefore (Wang et al., 2011) proposed a variant of HDP which they called *Online Hierarchical Dirichlet Process*. The Online-HDP is designed to analyze streams of data. It provides the flexibility of HDP and the speed of online variational Bayes. The *online* aspect of the Online-HDP is both a performance improvement and a variation in how models are updated. The improvement in performance is achieved by not having to crawl through the entire corpus repeatedly. Instead of having several passes with a fixed set of data, it was suggested by (Wang et al., 2011) the updates of the model can be optimized by iteratively choosing a random subset of the data, and then updating the variational parameters based on the current subset of data.

5 NEWS DETECTION EXPERIMENT

In the experiment part of this work, we utilize a set of *tweets* to train a topic model using the LDA and HDP models. We train both models by treating each *tweet* as a separate document and also by aggregating all *tweets* of a single user into one document. This gives us four different approaches to compare. Our motivation for training the models in two different ways using the same dataset is to see if we can counter the biggest problem with modeling *tweets* which is their short length. As this aggregation of *tweets* is an attempt to emulate the AT-model, we refer to the experiments using this dataset as LDA-AT and HDP-AT.

Before briefly describing the pre-processing steps and training of the models, we describe the data set and the data collection process. After that, we test the various methods described in the previous section on a set of test *tweets*, separate from the training *tweets*.

5.1 Data Set

For our experiments, we have fetched data from Twitter's own streaming API[3] and built a separate data set. Our training data set contains approximately 600,000 *tweets* from the New York (USA) area, collected in the period from November 26, 2014 to December 5, 2014.

Using Twitter's API it is possible to get an extensive set of metadata connected to each *tweet*. Most of this metadata is of no significance for topic modeling, and is stripped away. For the purposes of the experiment, the only things we keep are the screen name of the author and the content of the *tweet* itself.

[3] http://dev.twitter.com

Based on an idea from Meyer et al. (Meyer et al., 2011), we ignored any *tweet* containing non-ASCII characters. The rationale behind this idea is that there are several *tweets* containing nothing but emoticons, as well as several *tweets* written using non-ASCII characters (such as Arabic and Chinese). The danger of doing this is that a few relevant *tweets* might also be removed. We have nevertheless decided this is a fair tradeoff and negligable in order to find the majority of the news including *tweets*. If we were to include *tweets* made in foreign languages using a non-ASCII alphabet, the complexity of the analysis would dramatically increase. Moreover we are interested in presenting breaking news to a user who, presumably, knows English but not necessarily other languages. Another aspect of ignoring non-ASCII characters is that *tweets* containing unicode emoticons are mostly status updates or chatter, which can be safely ignored. Finally, if an actual news post gets filtered out due to including non-ASCII characters, chances are high that someone else will have posted about this same event without utilizing non-ASCII characters. All these points considered, we decided that it was a fair tradeoff to ignore all the *tweets* containing non-ASCII characters. In our data set around 30% of the *tweets* are ignored due to their inclusion of non-ASCII characters.

Other than removing *tweets* containing non-ASCII characters, we have replaced all the URLs with the word “LINK”. We kept all hashtags but we removed words that are not meaningful like combination of many words prepared as a hashtag but lacks the ’#’ sign. We also removed common stop words since they have no effect on analyzing the meaning of *tweets* by using topic modeling. We have not performed any other word processing on the *tweets* or stemming on the words. We also made a copy of the data set where all the *tweets* belonging to a user merged together.

Before performing the experiment, we trained the different algorithms on the dataset described above. We used the Python library Gensim[4] as it has embedded support for both LDA and HDP. For both models, the number of topics were set to 50. This number was chosen based on earlier research done by Hong and Davison (Hong and Davison, 2010) which showed that the best results were given if the number of topic were set to 50. There are some additional parameters that can be tweaked for the models, and we tried a few different settings. First we made the LDA model update in chunks of 1000. This means that the model is trained with 1000 *tweets* at a time. For the training set where each user’s messages are aggregated, this meant that the model was updated in chunks of 1000 unique users. By doing so in our experiments, we observed that this caused one topic to be “inflated”. Almost every single message we attempted to classify using a model trained in this fashion was classified into the same, highly general topic. By increasing the chunk size by a factor of ten to 10,000, this phenomenon disappeared.

Furthermore, we set the number of passes to perform with LDA to 10. This was chosen to ensure the initial training set converged well on topics. For HDP, we set the chunk size to 256. When doing online training (meaning that updating the model with real-time *tweets*), it is not possible to change these parameters. So the model is simply updated straight away with the provided corpus (new messages received in real-time). It is somewhat possible to adjust the chunk size with real-time messages by doing batch updates. The size of the batches will have to be balanced around not being too rare in addition to not being too small so that they skew the models. A compromise here from our trials seem to be update at about every 500-1000 new *tweets* that arrive.

After training the models, we used them on a set of 100 *tweets* collected in the same time period as the testing set. These *tweets* were new in that they were not part of the training set.

Additionally we used the models on portions of the New York Times annotated corpus (Sandhaus, 2008) to assess which topics were most frequently found in actual news articles. We did this by tallying up the most relevant topic for each of the articles in the corpus. Doing this gave us a list over the topics most related to the articles in the New York Times corpus. The “score” of a topic, then, is the number of articles in the New York Times corpus that topic was the best fit for. This was done as a part of filtering tweets concerning news and not as part of the experiment described below.

5.2 Experiment

To conduct our experiment we asked 7 people to participate for manual grading of our results. As mentioned above, we collected 100 *tweets* for our experiment. We utilized each of the trained models (LDA, LDA-AT, HDP, HDP-AT) to topic model these new *tweets*. The results of this topic modeling was distributed to our participants where they graded how well the topic assigned fit on a scale from 1-5. Here a score of 1 means the topic is not relevant, and 5 means it is a perfect fit.

As shown in table 1, some of the resulting topics are very generic and cover a broad set of terms. This is

[4] https://radimrehurek.com/gensim/

Table 1: An example of words in topic #32 and #1 found by LDA-AT.

Topic #32	
WORD	**PROB.**
good	0.0216
love	0.0207
time	0.0205
day	0.0188
today	0.0150
night	0.0124
great	0.0103
work	0.0099
life	0.0093
youre	0.0089

Topic #1	
WORD	**PROB.**
game	0.0649
played	0.0490
team	0.0354
win	0.0345
play	0.0253
football	0.0147
games	0.0145
ball	0.0114
pick	0.0112
points	0.0102

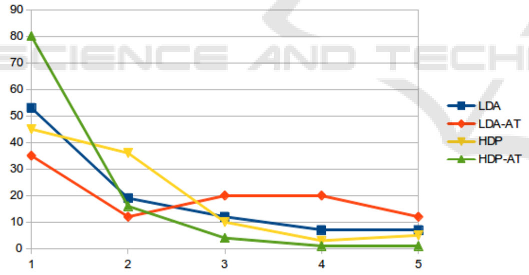

Figure 5: Results from modeling 100 tweets using LDA, LDA-AT, HDP and HDP-AT.

not surprising, as many *tweets* are about the everyday affairs of their authors. As we collected *tweets* from the New York area we expected an abundance of *tweets* about things occurring there.

Figure 5 shows the scores given to the categorization of each tweet in our test set by our test subjects. As is evident there is a clear difference in performance between the four methods. The two HDP variants, HDP-AT in particular performed rather poorly than HDP having an average score of 1.792, and HDP-AT having an average score of 1.178. The LDA variants performed much better compared to HDP methods. LDA had an average score of 2.000 where LDA-AT had an average score of 2.610 given by the participants of the experiment.

Table 2: Precision results given by the different topic modeling methods.

Method	**Precision**
LDA	0.287
LDA-AT	0.520
HDP	0.059
HDP-AT	0.198

To measure the results and give a final score of the different methods, we use precision. We set the threshold for a topic being relevant for a *tweet* at 3 out of 5. This means that every assignment with a score of 3 or higher, gets marked as relevant, while any assignment graded 2 or lower gets marked as not relevant. We calculate the precision by taking the number of relevant assignments and dividing them by the total number of assignments.

Using the scores given by the test participants and the definition above, we calculated the precision of each method. These scores can be found in table 2.

As mentioned previously, the sparsity and noisy nature of *tweets* make it difficult to get reasonable data out of them. This is especially valid when we strip them of stop words. After this process some *tweets* end up with a very low word count, and as such is not likely to get a suitable topic assigned. Furthermore, *tweets* in foreign languages are a challenge, and is not something we have taken into account in this work. Even when ignoring *tweets* containing non-ASCII characters, some languages (such as Spanish) still slip through.

As the results show, LDA-AT outperforms the other three models. The main rationale behind this is that by combining all *tweets* from a single author in the training set into a document (meaning the training set then contains one document per author), it becomes possible to somewhat counter the document length limitation inherent to *tweets*. As authors tend to stick to only a handful of topics, this should not skew the model in any meaningful way.

6 TOPIC MODELING COMBINED WITH NEWS INDEX

To be able to filter out what is news on Twitter, we utilized the New York Times annotated corpus, as described earlier. We modeled all the articles published by the New York Times in 2006 using the LDA-AT approach. We chose that approach, as our previous experiment had suggested it had the best performance on *tweets* of the four approaches tested. This gave us a

Table 3: Test results showing the number of relevant and non-relevant tweets with their precision values in different categories.

Experiment			
#Categories	**Relevant**	**Not rel.**	**Precision**
3 categories	7	247	0.028
2 categories	7	101	0.065
1 category	4	35	0.103

> *I'm on grand jury watch in the Eric Garner case. They are meeting and could decide today. Follow here and @NYTMetro for breaking updates*

Figure 6: An example of a news tweet found by using the topic modeling method combined with the news index.

handful of topics that were much more likely to be assigned to actual news articles than others. We then ran some fresh *tweets* through the model, and those who assigned to one of the top ranked topics are saved to see if they were actually news items. We did this three times; one with the top 3 topics, one with the top 2, and finally one with only the top topic. The results can be seen in table 3.

As is immediately clear, our approach for extracting news does not perform very well. The data set tested consisted of 3,454 tweets, meaning more than 90% of the total tweets were removed. Even so the best precision achieved was merely 0.103.

7 DISCUSSION

In this work we experimented different methods for detecting breaking news from Twitter streams. As a result of our comparison of different methods for topic modeling *tweets*, it seems using an LDA model coupled with aggregating all the *tweets* of a user, called LDA-AT, is the most effective approach. The largest limitation of using only topic modeling for news detection, is that it is sometimes difficult to know what a topic represents. Another pervasive limitation is that of Twitter's 140 characters limit on *tweets*.

Using the LDA-AT approach is a challenge when working with streams of real-time *tweets*. In a real-time setting the goal is to model *tweets* as they arrive. As our results show, combining all the *tweets* of a single user into one document is desired when performing this. In a real-time setting, however, one cannot afford to wait for *tweets* for an extended period of time to make sure that each user's document is large enough before updating the model. This is because, that would compromise the goal of topic modeling the tweets as soon as they arrive. One potential solution to this for news detection purposes would be to use a static topic model. On the other hand, this have the risk of the model getting outdated as popular topics on Twitter drift. Another potential solution for the real-time processing is to incrementally update the model in a set time, so as not to overly delay the topic modeling of the *tweets* themselves. This solution comes with another issue, however. The topic model we have used does not allow for terms to be added to the dictionary after the model has been initialized. This can be alleviated by using a hash map based dictionary, with the caveat that certain terms will share the same index and potentially lowers the precision of the model. The alternative is to keep the dictionary static. The danger of doing this is that over time certain terms that are not in the dictionary could become important to identify news.

However, the inherent limitation of 140 words per tweet is problematic for any statistical model that draws on word frequencies. Only a few content-relevant nominal phrases are included, and most of these 140 words tend to come from the stop word list. Even though the approach may be improved somewhat, the experiments seem to suggest that topic modeling is far from being effective in detecting breaking news on Twitter in the near future. Other and simpler techniques, like detecting clusters of tweets at particular locations at particular times, may be both computationally more efficient and quality-wise more precise.

On the other hand, topic modeling has some other advantages that may be interesting as part of a larger news aggregator service. Associating tweets with multiple topics, we can organize news tweets according to multiple dimensions for easier user inspection. We may also use the most prominent words of each topic representation as a short summary or title of a group of related tweets, making it unnecessary to check every tweet to understand the overall content.

8 CONCLUSION AND FUTURE WORK

In this paper, we have compared four different methods for topic modeling *tweets* as part of a breaking news detection system. As a result of our experiments LDA-AT outperforms the other three models. We also attempted to use the trained topic model in a practical manner to detect news. We used the New York Times annotated corpus to decide which topics were most likely to be assigned to news articles, before using those topics to filter a new data set. The results of this experiment show that the majority of tweets fetched using this method is non news, achieving a precision of only 0.103 in the best case.

Topic modeling itself is not likely to be sufficient for detecting breaking news from Twitter. The tweets are too short and too ambiguous to generate statistical models of the necessary precision. As a supplement to other techniques for news detection, they may however be useful, since they assume no knowledge of location, time or author.

From a news aggregator perspective topic modeling is interesting also for clustering and summarizing news content. Each tweet is associated with a number of relevant topics or clusters, and each topic is again described using a set of prominent word for that topic. In the future we intend to further explore the clustering abilities of topic modeling to improve the user experience of our news aggregator. It allows us to structure news content along several dimensions and use short labels to summarize sets of news stories.

REFERENCES

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Gulla, J. A., Fidjestøl, A. D., Su, X., and Castejon, H. (2014). Implicit user profiling in news recommender systems.

Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM.

Hu, M., Liu, S., Wei, F., Wu, Y., Stasko, J., and Ma, K.-L. (2012). Breaking news on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2751–2754. ACM.

Ingvaldsen, J. E., Gulla, J. A., and Özgöbek, Ö. (2015). User controlled news recommendations. In *Proceedings of the Joint Workshop on Interfaces and Human Decision Making for Recommender Systems colocated with ACM Conference on Recommender Systems (RecSys 2015)*.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

Mendoza, M., Poblete, B., and Castillo, C. (2010). Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM.

Meyer, B., Bryan, K., Santos, Y., and Kim, B. (2011). Twitterreporter: Breaking news detection and visualization through the geo-tagged twitter network. In *CATA*, pages 84–89.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., and Steyvers, M. (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems (TOIS)*, 28(1):4.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Sandhaus, E. (2008). The newyork times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476).

Titov, I. and McDonald, R. (2008). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120. ACM.

Wang, C., Paisley, J. W., and Blei, D. M. (2011). Online variational inference for the hierarchical dirichlet process. In *International Conference on Artificial Intelligence and Statistics*, pages 752–760.

Wang, X., Zhu, F., Jiang, J., and Li, S. (2013). Real time event detection in twitter. In Wang, J., Xiong, H., Ishikawa, Y., Xu, J., and Zhou, J., editors, *Web-Age Information Management*, volume 7923 of *Lecture Notes in Computer Science*, pages 502–513. Springer Berlin Heidelberg.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer.