A Hierarchical Topic Modelling Approach for Tweet Clustering

Bo Wang[1], Maria Liakata[1,2], Arkaitz Zubiaga[1], and Rob Procter[1,2]

[1] Department of Computer Science, University of Warwick, UK
[2] The Alan Turing Institute, London, UK
{bo.wang, m.liakata, a.zubiaga}@warwick.ac.uk

Abstract. While social media platforms such as Twitter can provide rich and up-to-date information for a wide range of applications, manually digesting such large volumes of data is difficult and costly. Therefore it is important to automatically infer coherent and discriminative topics from tweets. Conventional topic models and document clustering approaches fail to achieve good results due to the noisy and sparse nature of tweets. In this paper, we explore various ways of tackling this challenge and finally propose a two-stage hierarchical topic modelling system that is efficient and effective in alleviating the data sparsity problem. We present an extensive evaluation on two datasets, and report our proposed system achieving the best performance in both document clustering performance and topic coherence.

Keywords: tweet clustering, topic model, Twitter topic detection, social media

1 Introduction

In recent years social media platforms are increasingly being used as data sources to collect all kinds of updates posted by people. Updates that are of interest range from journalistic information that news practitioners can utilise for news gathering and reporting [25, 14], as well as opinions expressed by people towards a broad range of topics. While social media is a rich resource to shed light on public opinion and to track newsworthy stories ranging from political campaigns to terrorist attacks, it is often difficult for humans to keep track of all the relevant information provided the large volumes of data. Automatic identification of topics can help to produce a manageable list that is easier to digest for users, enabling for instance identification of real-world events among those topics.

In contrast to the well-studied task of Topic Detection and Tracking [2], which is concerned with topic detection from newswire articles, detecting topics in social media such as Twitter poses the challenges of dealing with unmoderated, user-generated content. This presents caveats such as inconsistent vocabulary across different users as well as the brevity of microposts that often lack sufficient context. As a consequence, traditional document clustering approaches using bag-of-words representation and topic models relying on word co-occurrence fall short of achieving competitive performance.

Recently a number of studies have employed various topic modelling approaches to tweets [30, 36, 38, 26], reporting mixed results and proving it to be a challenging task. In this work, we are motivated to effectively group tweets to a number of clusters, with each cluster representing a topic, story or event. Specifically, we propose a two-stage hierarchical topic modelling system shown in Figure 1, which: 1) uses a collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model (GSDMM) [38] for tweet clustering; 2) aggregates each tweet cluster to form a virtual document; 3) applies the second stage of topic modelling to the virtual documents but this time incorporates word embeddings as latent features (LFLDA) [26]. This not only alleviates the noisy nature of tweets but also generates meaningful and interpretable topics. Finally we conduct extensive evaluation on two datasets, using clustering evaluation metrics as well as topic model quality metrics. We compare our proposed approaches with other clustering-based methods and topic models, reporting the best scores in both clustering performance and topic coherence.

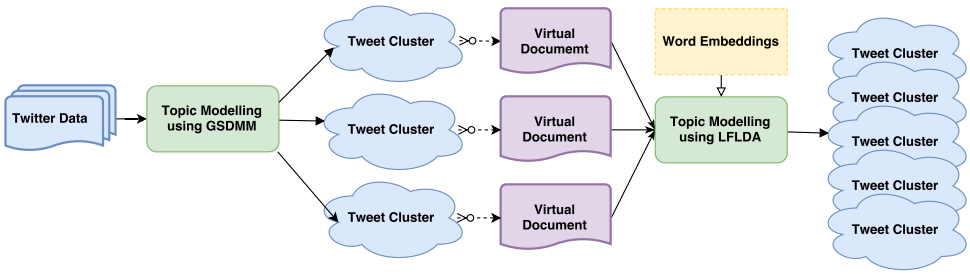

Fig. 1. Overview of the proposed topic modelling system

2 Related work

Conventional topic models such as Latent Dirichlet Allocation (LDA) [6] have shown great success in various Natural Language Processing (NLP) tasks for discovering the latent topics that occur in long and structured text documents. Due to the limited word co-occurrence information in short texts, conventional topic models perform much worse for social media microposts such as tweets as demonstrated by Rosa et al. [32]. In this section we review the recent developments on Twitter topic modelling and how to tackle the sparse and noisy nature of tweets.

Earlier studies try to utilise external knowledge such as Wikipedia [30] to improve topic modelling on short texts. This requires a large text corpus which may have a domain issue for the task at hand. Since then four approaches have

been studied in the literature to adapt conventional topic models for short texts such as tweets:

1) Directly model the generation of word co-occurrence pattern (i.e. biterms) as demonstrated by Yan et al. [37]. However, such word co-occurrence information is still limited to the 140 characters of each tweet.

2) Apply a document pooling strategy, to aggregate tweets to a number of virtual documents, based on authors [36], hashtags [21], conversation [3] or other metadata [11] such as timestamps and named entities. This strategy helps to overcome the limited context information in tweets, but pooling by such metadata can potentially have adverse effect on the subsequent topic modelling.

3) [27] proposed a simple topic model, named Dirichlet Multinomial Mixture (DMM) model, based on the assumption that each document is sampled from one single latent topic. The DMM model has since then been used in many Twitter topic modelling studies for alleviating the data sparsity problem and reported to give more coherent topics [40, 38, 31, 18], given that its underlying assumption is reasonable for short texts.

4) Complement topic models which use the global word collocation patterns in the same document/tweet, with word embeddings that exploit the local word collocation patterns within a context window. [26] extend LDA and DMM to incorporate word embeddings as latent features. Such latent feature component is integrated with its original topic-word Dirichlet multinomial component. [18] propose to incorporate word embeddings through the generalised *Pólya urn* model in topic inference. [12] propose to infer topics via document-level co-occurrence patterns of latent concepts instead of words themselves. All of these approaches aim to improve topic coherence by connecting semantically related words to overcome the short length of tweets.

In this paper, we present a comparative study on both topic modelling and document clustering approaches over two datasets, namely a first story detection corpus [29] and a large-scale event detection corpus covering over 500 events [20]. Our proposed two-stage topic modelling system adopts three of the four strategies mentioned above, achieving not only the best performance measured in document clustering metrics but also topic coherence for its generated topics.

3 Methodology

In recent years we have witnessed various topic modelling studies tackling the challenge of clustering tweets into topics using several different strategies, and yet it is still proven to be a difficult task to solve. Inspired by the two-stage online-offline approach in Twitter event detection studies [5, 39], we propose a two-stage hierarchical topic modelling system consisting of two state-of-the-art topic models, namely GSDMM [38] and LFLDA [26], with a tweet-pooling step streamlining the whole clustering process.

In the collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model [38] (GSDMM), the probability of a document belonging to a cluster is proportional to the cluster size and the frequency of each word of the

document in the cluster. More specifically after the initialisation step where documents are randomly assigned to K clusters, at each iteration it uses three count variables to record the information of each cluster: n_z^w which is the frequency of word w in cluster z, n_z which is the number of words in cluster z and m_z which is the number of documents in cluster z. Given its proven record on clustering tweets, we use GSDMM as the first stage of topic modelling and set K to be a very large number which allows GSDMM to automatically infer the final number of clusters.

As shown in Figure 1, we then assign every tweet to its corresponding cluster and aggregate each cluster to form a virtual document that consists of every tweet in that cluster. This pooling step is very similar to previous work [36, 21, 3], with the difference that it does not use any metadata which may not be available always (e.g. not every tweet mentions a hashtag or named entity).

Finally we apply the second stage of topic modelling to the previously generated virtual documents. Here we are motivated to take advantage of word embeddings [22] which have been shown to perform well in various NLP tasks, and combine it with topic models. [26] achieves this by replacing its topic-term multinomial distribution with a two-component mixture of a Dirichlet multinomial component and a word embedding component. We choose the better performing LFLDA model for our second-stage of topic modelling. Thus each tweet is assigned a topic with the highest topic proportion[1] given the virtual document cluster that it is in.

4 Datasets

We compare our proposed system with aforementioned approaches on two datasets, with different characteristics that help us generalise our results to different topic modelling tasks:

- A first story detection (FSD) corpus [29] collected from the beginning of July to mid-September 2011, containing 2204 tweets with each tweet annotated as one of 27 real-world stories such as "Death of Amy Winehouse" and "Terrorist attack in Delhi". It has some overlap of stories as well, e.g. four of the stories are related to the London riots in 2011, makes it also applicable to the task of sub-story detection.
- A large-scale event detection (ED) corpus [20], collected during October and November of 2012. Using Wikipedia and crowdsourcing as well as event detection methods [28, 1], it generated 150,000 tweets over 28 days covering more than 500 events. Each event label represents a specific topic or story line, e.g. "British prime minister David Cameron and Scottish first minister Alex Salmond agree a deal". After retrieving 78,138 tweets we decide to use the first five days of data for evaluation, resulting in five sets of *tweets/labels*: *3330/32, 2083/41, 6234/48, 2038/36 and 3468/43.*

[1] Topic proportion: the proportion of words in document d that are assigned to topic t or the topic probabilities of a document, i.e. $p(t|d)$

5 Evaluation

Experiments are conducted in two tasks. Moreover, document clustering metrics as well as topic model quality metrics are used for evaluation.

5.1 Experimental setup

Compared Methods: Both topic modelling and document clustering methods are evaluated. The topic modelling methods are:

- **OLDA** [10]: An online variational Bayes (VB) algorithm for LDA, based on online stochastic optimisation.
- **TOLDA** [16]: An online version of LDA specific for tracking trends on Twitter over time. Due to the limitation of the FSD corpus, this method is only evaluated in the event detection data [20].
- **GSDMM** [38]: A collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture (DMM) model, proven to work well for short texts.
- **LFTM** [26]: Consists of two models: **LFLDA** and LFDMM. We select the better performing LFLDA [12, 19] to evaluate, which LFLDA is an extension of LDA by incorporating word embeddings.
- **LCTM** [12]: A latent concept topic model, where each latent concept is a localised Gaussian distribution over the word embedding space.

For the above models we assign the topic with the highest topic proportion to each tweet.

As for document clustering baseline methods, we use the learnt topic proportion from the above topic models as feature for each tweet and apply a clustering algorithm, e.g. **OLDA+HC**. Additionally, we also evaluate a tweet clustering approach [35] that uses character-based tweet embeddings (i.e. Tweet2Vec [8]) and outperforms the winner [13] of the 2014 SNOW breaking news detection competition[2,3] which was defined as a topic detection task. This method was named as **Tweet2Vec+HC**. All document clustering baselines employ a hierarchical agglomerative clustering algorithm as it is proven to be effective in [35].

The same preprocessing steps are applied to all methods to reduce the noise level. This includes removing hashtag symbols, URL links, user mention symbols and punctuation as well as lower-casing and the tokenisation of each tweet.

Experimental settings: **GSDMM** infers the number of clusters automatically based on a pre-defined upper bound, we set this initial number to 100 (which is a large number comparing to the true number of clusters). For all other topic models including the ones in our proposed system we set the number of topics, $K = 100$, even if they are in the second stage of topic modelling. We use *GloVe*[4] word embedding representation for **LFTM** and **LCTM**.

[2] `http://www.snow-workshop.org/2017/challenge/`

[3] Their data is not evaluated due to its lack of annotated tweets.

[4] `https://nlp.stanford.edu/projects/glove/`

For **LFTM** we empirically set $\beta = 0.2$, $\lambda = 0.6$ for processing tweets; and $\beta = 0.1$, $\lambda = 0.6$ for virtual documents in the second stage of topic modelling. The number of latent concepts S in **LCTM** is set to 500. The number of iterations in **GSDMM** is set to 100. Other parameters are kept to their default settings.

For **Tweet2Vec+HC** we directly use the Tweet2Vec model from [35] trained using 88,148 tweets, also the same hierarchical clustering algorithm implementation from *fast-cluster* library [23]. Hierarchical clustering requires to choose a distance metric, linkage method and criterion in forming flat clusters. We evaluate the performance of different linkage methods and a wide range of distance metrics, using the Cophenetic Correlation Coefficient (CPCC) [34] and pick the best performing combination. The mean Silhouette Coefficient [33], a cluster validity index, was found to be the most effective among 30 validity indices for measuring the quality of the produced clusters [4]. To avoid using the ground truth labels, we select the optimal criterion and distance threshold according to the Silhouette score in a grid-search set-up. This way we make sure our comparisons are reasonable and unbiased.

5.2 Tweet clustering evaluation

With topic models, we can represent each tweet with its topic distribution $p(topic|tweet)$. Hence we can evaluate the performance of each topic model on a document clustering task, by using the topic proportion directly as the final cluster assignment or indirectly as feature representations for a further round of clustering or topic modelling. We then compare the resulting clusters to the true cluster labels in two datasets. Normalised Mutual Information (NMI) is widely used for measuring the overlap between the cluster assignments and the ground truth labels. It ranges from 0.0 (worst) to 1.0 (best). We select NMI as our clustering evaluation metric.

Table 1 presents the performance of the different methods on both datasets. Among the standalone topic models, GSDMM consistently outperforms other methods except for day-2 of the event detection (ED) corpus where it is beaten by OLDA by a small margin. OLDA showing surprisingly good performance across the board, credits to the online nature of its optimisation. The models that incorporate word embeddings, namely LFLDA and LCTM, show inconsistent performance over the two datasets. Different to what is reported in [12], we found that LCTM performs worse than LFLDA in half of the cases[5], potentially caused by the noisy nature of tweets and its adverse effect on constructing latent concepts. In general the two online models perform reasonably well for this task. As for Twitter Online LDA (TOLDA), interestingly we observe it performs worse than OLDA on the ED corpus, due to the large number of clusters it assigns to the tweets.

We observe mixed results by employing hierarchical clustering using topic proportions as features. In many cases it is showing to give almost equivalent

[5]We have also evaluated LCTM with number of concepts setting to 600 and 1000, however we observed little difference in the performance.

performance than using any topic model alone. This shows by simply using topic proportion as features for clustering is not a promising approach. We also observe by using Tweet2Vec neural embeddings with HC, it generates large number of clusters and thus very poor result.

Our two-stage topic modelling methods have shown to be rather effective in improving clustering performance, as only in 2 out of the 33 cases we have seen performance drop when comparing to either one of the topic models employed by the method (i.e. TOLDA+OLDA performs worse than OLDA at day-1 and day-2). This shows the promising result of using our proposed hierarchical topic modelling process with a pooling step. The proposed GSDMM+LFLDA proved to achieve consistent best performance over different datasets except at day-4 of the ED corpus it is beaten by GSDMM+OLDA.

Model	FSD		Day-1		Day-2		Day-3		Day-4		Day-5	
	N	**NMI**	***N***	**NMI**	***N***	**NMI**	***N***	**NMI**	***N***	**NMI**	***N***	**NMI**
OLDA	51	0.778	58	0.837	45	0.863	73	0.539	55	0.680	55	0.675
TOLDA			100	0.740	100	0.761	100	0.537	100	0.655	100	0.639
GSDMM	45	0.878	46	0.858	53	0.850	53	0.676	51	0.745	42	0.786
LFLDA	92	0.801	95	0.764	89	0.818	100	0.506	98	0.610	99	0.596
LCTM	93	0.721	94	0.726	83	0.804	100	0.512	99	0.632	97	0.617
OLDA+HC	42	0.799	39	0.828	40	0.859	64	0.529	45	0.684	49	0.669
TOLDA+HC			99	0.740	100	0.760	100	0.539	100	0.656	100	0.641
GSDMM+HC	45	0.878	46	0.859	53	0.851	53	0.677	51	0.745	94	0.771
LFLDA+HC	53	0.812	65	0.777	37	0.797	72	0.501	51	0.605	52	0.593
LCTM+HC	90	0.740	66	0.769	80	0.831	9	0.142	8	0.238	10	0.386
Tweet2Vec+HC	713	0.526	805	0.553	684	0.626	331	0.403	677	0.543	832	0.473
TOLDA+OLDA			32	0.819	34	0.847	35	0.613	38	0.696	35	0.755
TOLDA+LFLDA			48	0.814	46	0.845	40	0.577	41	0.706	35	0.718
TOLDA+LCTM			45	0.812	35	0.856	41	0.544	43	0.692	40	0.692
GSDMM+OLDA	36	0.891	26	0.870	34	0.872	38	0.694	25	**0.816**	26	0.793
GSDMM+LFLDA	30	**0.926**	28	**0.871**	29	**0.882**	34	**0.695**	27	0.773	22	**0.812**
GSDMM+LCTM	32	0.912	41	0.861	39	0.860	43	0.663	39	0.765	35	0.789

Table 1. Document clustering performance (NMI only) on both datasets

5.3 Topic coherence evaluation

Here we examine the quality of our hierarchical topic modelling system by the topic coherence metric. Such metric measures to what extent the top topic words, or the words that have high probability in each topic are semantically coherent [7]. This includes using word intrusion [7], Pointwise Mutual Information (PMI) [24] and Normalised PMI (NPMI) [17]. We adopt the word embedding-based topic coherence metric, proposed in [9], which is shown to have a high agreement with humans and are more robust than the PMI-based metrics for tweets. In this paper we use two pre-trained word embedding models learnt from Twitter data[6], resulting in two metrics G-T-WE (GloVe) and W-T-WE (Word2Vec). We also

[6]The GloVe model was trained using 2 billion tweets while the Word2Vec model was trained on 5 million tweets using the skip-gram algorithm.

adopt the approach in [15], computing coherence for top-5/10/15/20 words and then take the mean over the 4 values.

For the ED corpus, we average all the results over the 5-day period for each model. As shown in Table 2, GSDMM+LFLDA achieves the best topic coherence in 3 out of 4 cases, with TOLDA+OLDA outperforming the others for W-T-WE on the ED data. When we compare the the two-stage topic modelling approach (i.e. TOLDA+* or GSDMM+*) to its respective topic model used in the first stage (i.e. TOLDA or GSDMM), we observe in 10 out of 12 cases its topic coherence has improved. Though our results for coherence are not perfect, it is demonstrated the usefulness of aggregating first round tweet clusters into virtual documents without the use of any metadata and then performing second round of topic modelling. As a result it is able to create not only more discriminative but also more coherent clusters.

Model	**Topic Coherence**			
	FSD		Event Detection	
	G-T-WE	W-T-WE	G-T-WE	W-T-WE
OLDA	0.217	0.123	0.302	0.135
TOLDA			0.329	0.141
GSDMM	0.277	0.121	0.363	0.132
TOLDA+OLDA			0.349	**0.154**
TOLDA+LFLDA			0.371	0.137
GSDMM+OLDA	0.282	0.142	0.349	0.150
GSDMM+LFLDA	**0.315**	**0.144**	**0.385**	0.142

Table 2. Averaged topic coherence for both corpora

5.4 Qualitative evaluation of topics

We also present a set of randomly selected example topics generated by the proposed system, GSDMM+LFLDA, on both data sets. Due to the limited space, these example topics are shown in Table 3 and Table 4 of the Appendix.

6 Conclusions and future work

Inferring topics in tweets is hard due to the short and noisy nature of tweets. In this paper we proposed a two-stage hierarchical topic modelling system, named GSDMM+LFLDA, that leverages a state-of-the-art Twitter topic model, a topic model with word embeddings incorporated and a tweet pooling step without the use of metadata in any form. We performed extensive experiments on two Twitter corpora. The experimental results show our proposed approach outperforms other clustering-based methods and topic models, in both clustering performance and topic coherence.

For future work, we plan to evaluate our system in tracking the same set of topics across adjacent time intervals, which is a different task to document clustering and topic detection.

Acknowledgments

This work is partly supported by The Alan Turing Institute. We would also like to thank Anjie Fang, Dat Quoc Nguyen, Jey Han Lau, Svitlana Vakulenko and Weihua Hu for answering questions regarding their work, respectively.

Appendix

We present a set of randomly selected example topics generated by GSDMM+LFLDA, on both the first story detection (FSD) corpus and the first day of the event detection (ED) corpus, as seen in Table 3 and Table 4. Each detected topic is presented with its top-10 topic words, and is matched with the corresponding topic description or story from the ground truth (given by the creators of these data sets), as well as a sample tweet retrieved using the topic keywords.

As shown in Table 3 and Table 4, words in obtained topics are mostly coherent and well aligned with a ground-truth topic description. We can also discover more useful information with regard to the corresponding real-world story, by simply looking at its topic words. For example, in the first topic of Table 3 we see the Twittersphere has mentioned 'Amy Winehouse' and 'death' along with the word 'drug'. This information may have been missed if one only chooses to read a set of randomly sampled tweets mentioning 'Amy Winehouse'.

Detected topic	**Corresponding topic description**	**Sample tweet**
amy winehouse rip amywinehouse die dead sad dy talent drug	Death of Amy Winehouse.	jesus, amy winehouse found dead. v sad #winehouse
tottenham riot police news fire shoot car london north thur	Riots break out in Tottenham.	RT @itv_news: Police cars set on fire in Tottenham, north London, after riots connected to the shooting of a young man by police on Thur ...
mars water nasa flow found evidence may scientist saltwater liquid	NASA announces discovery of water on Mars.	RT @CalebHowe: NASA reporting live right now that they have circumstantial evidence for flowing, liquid water on Mars.
house debt bill pass us vote ceiling the representatives raise	US increases debt ceiling.	RT @politico: On Monday evening the House passed a bill to raise the debt ceiling, 269 to 161.
delhi high blast court outside injured explosion attack kill bomb	Terrorist attack in Delhi.	Bomb Blast outside of High Court Delhi just few minutes ago. http://t.co/MejKWlC
pipeline fire kenya least kenyans people gasoline kill dead lunga	Petrol pipeline explosion in Kenya	RT @AKenyanGirl: RT @CapitalFM_kenya: Dozens suffer burns in Kenya #Pipeline fire in Lunga Lunga, Nairobi. Firefighters battling inferno ...

Table 3. Example topics detected on FSD corpus

Detected topic	Corresponding topic description	Sample tweet
merkel angela greece visit athens merkels greek chancellor protests protest	An estimated 25,000 protest in Athens as German Chancellor Angela Merkel visits Greece.	thousands protest merkel s greece visit http://t.co/sXGTX3jE
syrian plane turkey passenger turkish land ankara force syria intercepts	A Syrian passenger plane is forced by Turkish fighter jets to land in Ankara due to the allegations of carrying weapons.	BreakingNews: Turkish fighter jets force Syrian passenger plane to land in Ankara: Anadolu Agency
malala yousafzai taliban activist pakistan shot girl attack bullet shooting	Malala Yousafzai, a 14 year old activist for women's education rights is shot by Taliban gunmen in the Swat Valley.	Taliban Says It Shot Pakistani Teen, Malala Yousafzai, For Advocating Girls Rights... http://t.co/EjFR5in4
lenovo hp pc top market battle spot computerworld gartner shipments	HP and Lenovo battle for top spot in PC market of Computerworld.	HP, Lenovo battle for top spot in PC market - Computerworld http://t.co/zwzPdN8Q #googlenews
merger eads bae systems aerospace plans talks cancel defence firms	BAE and EADS announce their merger talks are cancelled over political disagreements.	BAE-EADS merger plans are 'off': Aerospace and defence firms BAE and EADS have cancelled their planned merger, t... http://t.co/UYFOiysX
pussy riot court appeal moscow member one freed russian punk	A court in Moscow, Russia, frees one of the three Pussy Riot members at an appeal hearing.	One Pussy Riot Member Freed by Moscow Court — News — The Moscow Times http://t.co/m60lwaWU #FreePussyRiot

Table 4. Example topics detected on ED corpus - day one

References

1. Aggarwal, C.C., Subbian, K.: Event detection in social streams. In: Proceedings of the 2012 SIAM international conference on data mining. pp. 624–635. SIAM (2012)
2. Allan, J.: Topic detection and tracking: event-based information organization, vol. 12. Springer Science & Business Media (2012)
3. Alvarez-Melis, D., Saveski, M.: Topic modeling in twitter: Aggregating tweets by conversations. In: ICWSM. pp. 519–522 (2016)
4. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., Perona, I.: An extensive comparative study of cluster validity indices. Pattern Recognition 46(1), 243–256 (2013)
5. Becker, H., Naaman, M., Gravano, L.: Beyond trending topics: Real-world event identification on twitter. ICWSM 11(2011), 438–441 (2011)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
7. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Advances in neural information processing systems. pp. 288–296 (2009)
8. Dhingra, B., Zhou, Z., Fitzpatrick, D., Muehl, M., Cohen, W.W.: Tweet2vec: Character-based distributed representations for social media. In: The 54th Annual Meeting of the Association for Computational Linguistics. p. 269 (2016)

9. Fang, A., Macdonald, C., Ounis, I., Habel, P.: Using word embedding to evaluate the coherence of topics from twitter data. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 1057–1060. ACM (2016)
10. Hoffman, M., Bach, F.R., Blei, D.M.: Online learning for latent dirichlet allocation. In: advances in neural information processing systems. pp. 856–864 (2010)
11. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the first workshop on social media analytics. pp. 80–88. ACM (2010)
12. Hu, W., Tsujii, J.: A latent concept topic model for robust topic inference using word embeddings. In: The 54th Annual Meeting of the Association for Computational Linguistics. p. 380 (2016)
13. Ifrim, G., Shi, B., Brigadir, I.: Event detection in twitter using aggressive filtering and hierarchical tweet clustering. In: Second Workshop on Social News on the Web (SNOW), Seoul, Korea, 8 April 2014. ACM (2014)
14. Jordaan, M.: Poke me, i'm a journalist: The impact of facebook and twitter on newsroom routines and cultures at two south african weeklies. Ecquid Novi: African Journalism Studies 34(1), 21–35 (2013)
15. Lau, J.H., Baldwin, T.: The sensitivity of topic coherence evaluation to topic cardinality. In: Proceedings of NAACL-HLT. pp. 483–487 (2016)
16. Lau, J.H., Collier, N., Baldwin, T.: On-line trend analysis with topic models:\# twitter trends detection topic model online. In: COLING. pp. 1519–1534 (2012)
17. Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: EACL. pp. 530–539 (2014)
18. Li, C., Wang, H., Zhang, Z., Sun, A., Ma, Z.: Topic modeling for short texts with auxiliary word embeddings. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 165–174. ACM (2016)
19. Li, S., Chua, T.S., Zhu, J., Miao, C.: Generative topic embedding: a continuous representation of documents. In: Proceedings of The 54th Annual Meeting of the Association for Computational Linguistics (ACL) (2016)
20. McMinn, A.J., Moshfeghi, Y., Jose, J.M.: Building a large-scale corpus for evaluating event detection on twitter. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management. pp. 409–418. ACM (2013)
21. Mehrotra, R., Sanner, S., Buntine, W., Xie, L.: Improving lda topic models for microblogs via tweet pooling and automatic labeling. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. pp. 889–892. ACM (2013)
22. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
23. Müllner, D., et al.: fastcluster: Fast hierarchical, agglomerative clustering routines for r and python. Journal of Statistical Software 53(9), 1–18 (2013)
24. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 100–108. Association for Computational Linguistics (2010)
25. Newman, N.: The rise of social media and its impact on mainstream journalism (2009)
26. Nguyen, D.Q., Billingsley, R., Du, L., Johnson, M.: Improving topic models with latent feature word representations. Transactions of the Association for Computational Linguistics 3, 299–313 (2015)

27. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using em. Machine learning 39(2), 103–134 (2000)
28. Petrović, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to twitter. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 181–189. Association for Computational Linguistics (2010)
29. Petrović, S., Osborne, M., Lavrenko, V.: Using paraphrases for improving first story detection in news and twitter. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 338–346. Association for Computational Linguistics (2012)
30. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th International Conference on World Wide Web. pp. 91–100. ACM (2008)
31. Quan, X., Kit, C., Ge, Y., Pan, S.J.: Short and sparse text topic modeling via self-aggregation. In: IJCAI. pp. 2270–2276 (2015)
32. Rosa, K.D., Shah, R., Lin, B., Gershman, A., Frederking, R.: Topical clustering of tweets. Proceedings of the ACM SIGIR: SWSM (2011)
33. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 20, 53–65 (1987)
34. Sokal, R.R., Rohlf, F.J.: The comparison of dendrograms by objective methods. Taxon pp. 33–40 (1962)
35. Vakulenko, S., Nixon, L., Lupu, M.: Character-based neural embeddings for tweet clustering. SocialNLP 2017 p. 36 (2017)
36. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. pp. 261–270. ACM (2010)
37. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: Proceedings of the 22nd International Conference on World Wide Web. pp. 1445–1456. ACM (2013)
38. Yin, J., Wang, J.: A dirichlet multinomial mixture model-based approach for short text clustering. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 233–242. ACM (2014)
39. Yin, J.: Clustering microtext streams for event identification. In: IJCNLP. pp. 719–725 (2013)
40. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: European Conference on Information Retrieval. pp. 338–349. Springer (2011)