Improve topic modeling algorithms based on Twitter hashtags

Hayder M Alash¹, Ghaidaa A Al-Sultany²

¹Department of Software, College of Information Technology, University of Babylon, Babel, Iraq.
²Department of Information Network, College of Information Technology, University of Babylon, Babel, Iraq.

Corresponding author’s e-mail: hayder_albdulhameed@uobabylon.edu.iq, ghaidaa.almulla@it-net.uobabylon.edu.iq

Abstract. Today with increase using social media, a lot of researchers have interested in topic extraction from Twitter. Twitter is an unstructured short text and messy that it is critical to find topics from tweets. While topic modeling algorithms such as Latent semantic analysis (LSA) and Latent Dirichlet Allocation (LDA) are originally designed to derive topics from large documents such as articles, and books. They are often less efficient when applied to short text content like Twitter. Luckily, Twitter has many features that represent the interaction between users. Tweets have rich user-generated hashtags as keywords. In this paper, we exploit the hashtags feature to improve topics learned from Twitter content without modifying the basic topic model of LSA and LDA. Users who share the same hashtag at most discuss the same topic. We compare the performance of the two methods (LSA and LDA) using the topic coherence (with and without hashtags). The experiment result on the Twitter dataset showed that LSA has better coherence score with hashtags than that do not incorporate hashtags. In contrast, our experiments show that the LDA has a better coherence score without incorporating hashtags. Finally, LDA has a better coherence score than LSA and the best coherence result obtained from the LDA method was (0.6047) and the LSA method was (0.4744) but the number of topics in LDA was higher than LSA. Thus, LDA may cause the same tweets to discuss the same subject set into different clustering.

Keywords: Topic Derivation, Latent semantic analysis (LSA), Latent Dirichlet Allocation (LDA), Twitter, Hashtag.

1. Introduction

In recent years, social media such as Facebook and Twitter are widely used worldwide. Twitter is the largest social media network used to write and read the short text by people about anything in life. Therefore, social media has become a place where people are sharing their opinions and real feeling such as a special event, product, disaster, accident, and markets. These raise need to analyze what people talk about in social media by finding topics under discussion.

There are a variety of Topic model methods used to analyze the texts in large documents such as LDA [1], LSA [2], and Probabilistic latent semantic analysis (PLSA) [3]. Topic modeling is a technique for automatic classification of documents (unsupervised method) and understands a large set of
information in any large group of documents (corpus), where the frequency of terms overlapping and co-occurrences between terms can be high. Topics modeling algorithms are used to derive topics from the document in the corpus and identifying the numbers of topics in each document and find the dominant topic by using the important words for each topic. In contrast, Twitter is short text length (post) which includes many misspelling words, abbreviations, emoticons, and non-conventional syntax. Therefore, deriving topics from the Twitter short text (tweets) is a big challenge. The number of words to build frequency of co-occurrences between terms is low.

Most of the researches is focused only on the content of the text to derive a topic from social media without including other social network features. The social network provides many features that represent an interaction between users. On Twitter, replies, mentions, retweets, and hashtags represent social network features. Users can reply to another user post using reply feature. Mention is usually a Tweet that contains another username anywhere into text posted by a user about a particular topic. Retweet is a feature used by users to share others Tweets and can share own Tweets with all followers. A hashtag is a special symbol “#” starting word or phrase creates a hashtag on a user's tweet. Recently, researchers have found incorporate text with social features that can increase cohesiveness of topics extracted by topic modeling algorithms in a social media [4].

In this paper, we proposed to explore hashtags to improve the topic modeling performance, which uses the contents of tweets and specific Hashtags to perform topic derivation. The tweets have the same hashtags that have highly discussed the same topics. We evaluate this hashtag content technique using a coherence score. The Two methods (LSA and LDA) have been evaluated with the hashtag content model and the performance of the two methods has been compared using the topic coherence.

The rest of the paper was organized as follows section 2 presents a survey about the related works. Section 3 includes a general overview about topic modeling. Section 4 presents Methodology, and Section 5 shows experiments result. Finally, in Section 6 presents conclusion and future work.

2. Related Works

Recently, several researchers have focused on discovering the latent topics from text using topic models. Many existing topic models have been used for document analysis, such as LDA [1], LSA [2], and PLSI [3]. Topic model approaches have been successful derive topics for the long document in the traditional way. However, these approaches suffer when used to discover latent topics in social media such as Twitter. The Twitter environment has raised many challenges. A tweet is a very short text and includes many misspelling words, irrelevant characters, emoticons, and non-conventional syntax. This can lead to very low numbers of co-occurrence of terms within Tweets collection.

Several research tried to extracting latent topics from social media which are still based on the original methods (such as LDA, and LSA) used to derive the topic on the long document. A variety of methods have been proposed to improve the performance of finding topics in short texts. Some studies have been proposed to exploit social media features such as a hashtag, mention, and external resources to improve the coherence of the topics.

Many of the research still focuses on the content of the text. They have exploited external resources (such as Wikipedia) to add to content. [5] evaluated the performance of three topic modeling algorithms LDA, LSA, Non-Negative Matrix Factorization (NMF). They concluded that the LSA topic modeling algorithm gets high performance for short text (sentence) classification while the LDA algorithm gets high performance for large text (document) clustering. [6] Used Wikipedia article titles to identify the topic related to sets of posts on Twitter. They calculated words importance in tweets corpus and Wikipedia by TF_IDF then link Wikipedia data to identify topics in Twitter data using word similarity. [7] proposed an approach to attach descriptions from Wikipedia for each tweet. It converted short posts into long texts which complete the contextual information. [8] proposed a new pooling technique for tweets by aggregating tweets occurring in the same user-to-user interaction into longer documents. This technique showed better clustering quality and retrieval topic for the document. [9] proposed a new model to classify short text named sentimental bi-term topic model (SBTM) based on bi-term topic model which extended the short text by adding relative words. They exploited biterm topic model to
determine the topic probability distribution of documents and biterm, then finding similarity between two documents.

Some studies proposed exploited the word embedding technique to improve topic modeling in short text and social media. [10] designed new topic models GPU-DMM that enrich topic modeling for short texts by auxiliary word embedding from external documents. GPU-DMM improved the similarity between two short texts using semantically related words which rarely classify to the same topic. [11] proposed new model learn the distributed representations of words and entities by exploiting the latent semantic model. [12] extended LDA to derive the topic proportion of short text by combining the Incremental Biterm Topic Model (IBTM) to get better topics. [13] presented a new model to deal with short text called a word co-occurrence network-based model (WNTM). The model succeeded to enhance the semantic density of data without adding more time and complexity by using a sliding window to build a word network. In their work, they considered words occurring in the same window closely connected and high probability occur in the same document. [14] proposed aggregate similar tweets into large documents by applying several pooling techniques. Their method used hashtags or authors to aggregate tweets which showed improvement in topic coherence. [15] compared between different topic clustering techniques. They found the word2vec topic model algorithm worked better than another clustering in the online social network (OSN).

Recently, research has taken advantage of social media features to improve topic derivation in short text like Twitter. [16] proposed combining social hashtag graph feature and short text information which extend each text by associated hashtags to get whole semantic features. [17] proposed combines the social attributes and semantic features to increase the user query by utilizing an external search engine. [18] This paper proposed a new pooling technique by incorporating a community detection method to improve the topic derivation without making any changes to the LDA topic model. This approach aggregate user text content who shares topics and relation.

Our proposed method differs from previous works in that it exploits hashtags feature to improve topics derived from short text. The hashtag feature incorporates with tweets (text) to improve topics modeling without modifying the basic model of LSA and LDA.

3. Topic Modeling

Topic modeling is a statistical technique for discovering latent topics in a set of documents. It uses a text mining tool for discovering Latent semantic topics in a text body [14]. The document usually contains a mix of topics and that each topic consists of a set of words and can obtain the topics. Topic models work usually by group documents (tweets) that have common similar words and appear in a similar set of documents. They cluster documents based on their dominant topics that each topic consists of a list of important keywords [15]. Several kinds of research have studied topic modeling algorithms to derive topics. For example, latent semantic analysis (LSA) is one of the oldest technique exploits a set of terms and documents relationship using a term-document matrix. LSA uses the singular value decomposition (SVD) method to reduce the term-document matrix into k topics. Another topic model technique is Latent Dirichlet Allocation (LDA) which represented as the state-of-the-art approach in topic algorithm. LSA and LDA are models originally designed to derive topics from long documents, which can contain many topics. When applied to short text content like Twitter, They have a poor performer. Therefore, we proposed using the hashtag feature to improve topics learned from Twitter content without modifying the original topic model of LSA and LDA.

4. Methodology

Our approach start by discussing collect a dataset from Twitter API. Then, several preprocessing is applied to clean tweet text (post) from irrelevant words. After that, the cleaned tweets converted to features such as unigram and TF_IDF. Extraction features have passed to topic model algorithms for deriving topics. Finally, the results of our experiments were compared using the coherence score. Figure 1 represents the workflow of our proposal.
4.1. Twitter Dataset

Twitter provides an API stream platform to collect real-time tweets. Several of the studies used the API stream as a source of data. Therefore, we have used an API stream to collect tweets data using a set of predefined keywords and hashtags related to it. Tweets have collected from Twitter content of tweets (text) and several features (post time, user name, hashtag, retweet). The hashtag used as part of our tweet to improve train our model. English tweets have collected only in our experiments.

![Flowchart](image)

**Figure 1.** Workflow of our method

4.2. Preprocessing Stage

Twitter was a very unstructured short text and messy which is difficult to find topics from a set of tweets using topic modeling algorithms due to low terms(words) co-occurrence frequency. Therefore, we need to clean the collected data (Tweets) from Twitter API by converting the unstructured text in tweets into more acceptable readable structured for topic modeling. The reason for preprocessing step is to improve the accuracy of topic extraction. Tweets have many misspellings and meaningless words, symbols, irrelevant characters, and emoticons. Also, it contains many of stopwords, prepositions, punctuation, etc. in our research, we used the following text cleaning steps:

- Remove Mention from tweets which begin with '@' symbol and follow by another username anywhere in the text.
- Text normalization by removing all words less 3 which would not give an important meaning in the topic model, removing signs, removing all numbers, removing symbols, removing punctuations which didn’t contain any meaning in the text, and convert letters to Lowercase.
- Removing Web URL and Emails from the tweets.
- Removing distracting single quotes.
- Removing "Emoji" and slags words.
- Remove newline characters.
- splitting the text into small pieces called (tokens).
- Removing frequent words and stopwords e.g." my"," it", etc. which would not give specific semantic.

People use hashtags on Twitter which begin with a # symbol followed by a relevant keyword or phrase. Hashtag showed very important content in Tweets to categorize those Tweets and help them appear very easily in search. Therefore, our proposed method has exploited Hashtags content as a part of the tweet instead of deleting them, only '#' symbols removed from tweets' text. People have common
Hashtag may share the same topics which help topic model algorithms to clustering tweets. See algorithm 1.

### Algorithm 1

**Input:** Tweets  
**Output:** Extended tweets  

1. for all tweet in Tweets  
2. Tokenize(tweet)  
3. Preprocessing (tweet)  
4. Remove “#” from Hashtags  
5. Extract Hashtags followed relevant keyword or phrase  
6. Extend tweet by Hashtags

### Algorithm 2

**Input:** Extended tweets with Hashtags  
**Output:** Topics  

**Step 1:** For all ET in Extended Tweets  
\[ \text{TF-IDF} = \text{TF-IDF (ET)} \]  

**Step 2:** Select Number of k topics  
**Step 3:** For each word in TFIDF  
Find Topics in Extended Tweets

4.3. **Features Extraction**  
After the preprocessing stage, each text (tweet) was converted into an appropriate format (features vector) for topic models. One of these features is a bag of words matrix (Bow) where each document (tweet) is represented by the number of times a word found in a tweet. Another feature is Term Frequency - Inverse Document Frequency (TF_IDF) which gives higher weights for words occur frequently in the tweet but less often over the whole Tweets. TF_IDF is calculated using a mathematical equation below:

\[ TF_{IDF_{i,j}} = tf_{i,j} \times idf_{i}(\log \left( \frac{N}{df_i} \right)) \]  

(1)

Where \( tf_{i,j} \) refers to the number of occurrences that term \( i \) occurs in document \( j \), \( idf_i \) is the logarithm of dividing the number of tweets \( N \) by the number of tweets containing term \( i \). TF_IDF technique filters out common words and enables improve clustering models.

4.4. **Topic Modeling**  
The converted tweets into TF_IDF feature were used by LSA and LDA for topic modeling (see Algorithm 2.). The hashtags incorporate with tweets were tested in two models to check how the topics will affect by this proposal. In this paper, two techniques were compared as followings:

4.4.1. **Latent semantic analysis (LSA)**  
LSA [2] is an unsupervised approach to extract the latent topics from a set of documents. It was introduced to improve information retrieval by considering related content instead of depending on matching similar words and reduce the dimensionality of a matrix using Singular value decomposition (SVD). The LSA method deals nicely with the synonymy which can be described as various words or phrases that share the same meaning. LSA represents text as a set of term-document matrix and use Singular value decomposition (SVD) to decompose this matrix to lower dimensional and extract a latent k topics. SVD used to reduce matrix dimensionality which represents a semantic matrix to capture the relationships between words and documents. SVD is a square matrix which decomposed term-document matrix into three matrix multiplications as follows:

\[ X = T \times S \times D^T \]  

(2)

Where \( X \) is termed documents matrix with size the number of terms product the number of a documents. \( T \) and \( D^T \) represent the rank k reduce matrix and \( S \) is the singular values by diagonal
matrix. SVD allows a simple design to achieve the best approximate fit using smaller matrices by only taking the largest k singular values in S which are ordered by size.

4.4.2. A Latent Dirichlet Allocation (LDA).

LDA [1] is a very popular probabilistic model based on the Bayesian topic model. Each document has a random mixture of topics where each topic consists of a latent distribution of words. LDA automatically discovers the Latent topics from a set of documents. LDA process the document and words as follows the generative process:

- For each topic \( t \) have a word distribution \( \varphi_t \), choose \( \varphi_t \sim \text{Dir}(\beta) \)
- For each document \( d \) have a topic distribution \( \theta_d \), choose \( \theta_d \sim \text{Dir}(\alpha) \)
- For each word \( w_i \) in current document \( d \):
  - Select the topic \( t_i \sim \text{Multinomial}(\theta_d) \)
  - Select a word \( w_i \sim \text{Multinomial}(\varphi_{t_i}) \)

\[ 
\text{coherence}(V) = \sum_{(vi,vj) \in V} \text{score}(vi,vj,\varepsilon) 
\]

\( V \) represents a collection of the word describing the topic and \( \varepsilon \) represents a smoothing factor and guarantees score will return real numbers.

5. Experimental Results

Our experiments were tested on Twitter short text datasets to evaluate the performance of our method. Tweets were collected from Twitter using keywords and hashtags related to (Education, Car, Finance, and Entertainment). The Twitter dataset is extracted to get Tweets information including texts, hashtags, mentions. It consists 9159 tweets that contained hashtags in 1286 tweets. Texts and hashtags are used to input into LSA and LDA topic model algorithms.
Next, Tweets (text) are preprocessed by applying all processes mentioned in our method. Tweets cleaned and converted from unstructured text into suitable structured for topic modeling. Text cleaned by removing all words less than three characters which would not have any important meaning, removing all numbers, removing symbols, removing signs, removing punctuations and convert letters to Lowercase ...etc. For Hashtags, the text content keep as a part of the tweet which will extend the text to improve the topic modeling performance and the '#' symbol removed.

After preprocessing, the text data converted into proper topic model structures using Term Frequency - Inverse Document Frequency (TF_IDF) which helps to identify the important words into an entire corpus. Two topic modeling algorithms (LSA and LDA) with hashtags considered to find latent topics in the Twitter short text. Each topic model algorithm executes for a particular number k depends on the dataset. The coherence score is used to calculate coherent topics for each model.

**Table 1.** Coherence Score for LSA and LDA using our method

| Topic No. | LSA with Hashtags | LSA without Hashtags | LDA with Hashtags | LDA without Hashtags |
|-----------|-------------------|----------------------|-------------------|----------------------|
| 2         | 0.3229            | 0.4269               | 0.285             | 0.357                |
| 4         | 0.4744            | 0.385                | 0.3788            | 0.3279               |
| 6         | 0.4449            | 0.3679               | 0.4304            | 0.4368               |
| 8         | 0.3329            | 0.278                | 0.4856            | 0.5077               |
| 10        | 0.3955            | 0.3281               | 0.5047            | 0.5716               |
| 12        | 0.3229            | 0.3124               | 0.5944            | 0.5574               |
| 14        | 0.2615            | 0.2936               | 0.5867            | 0.6047               |
| 16        | 0.3384            | 0.2764               | 0.6036            | 0.6023               |
| 18        | 0.3136            | 0.3041               | 0.6027            | 0.6004               |

**Figure 3.** These LSA and LDA coherence Score (with and without) Hashtags

Several experiments have conducted on different setups for two methods. The number of the topics was used between k=2 and 20 and the coherence score was calculated for each method with k topics. LSA and LDA clustering methods applied to the TF_IDF vectors build from Twitter Dataset. The results of LSA and LDA methods with hashtags and without hashtags are listed in Table 1 and Figure 3. The best number of topics for LSA with hashtags is a Four with coherence score 0.4744 while The best number of topics for LSA without hashtags is two with coherence score 0.4269. For the LDA method, The best number of topics with hashtags is 16 topics with coherence score 0.6036 while The best number of topics with hashtags is 14 topics with coherence score 0.6047. A higher coherence score indicates to a topic is easier to understand the word distribution about what subjects would belong. The LDA Coherence score value has increased with the increasing of the number of topics. The result shows that
LSA incorporates hashtags that performs better than LSA without hashtags on most k topic numbers. Using hashtags feature with topic model methods helps to improve the coherence score value in our LSA result, while LDA performs better without incorporate hashtags. LDA has a higher coherence score value than LSA due to the Dirichlet process trying to find more topics in the text. In contrast, LSA is a semantic structure method that tries to collect words to share the same subject into the same topics. Therefore, the number of k topics in LSA is less than LDA due to collect all tweets share the same topics into one general topic. Thus, in the result, LSA coherence score values are less than LDA.

6. Conclusion

In this work, we proposed a method that incorporates hashtags feature with topic model algorithms in Twitter social media. Twitter is unstructured and short texts that it becomes a challenge to find topics from tweets. Several experiments of our method have conducted on a Twitter dataset. The results show the importance of using the hashtags feature with topic model methods to improve the coherence score in LSA. LSA with hashtags has outperformed than that the ones who do not incorporate hashtags. In this experiment, a comparison between LSA and LDA topic modeling is studied that LDA has a higher coherence score value than LSA because of LDA structure that tries to extract all topics in the text. While LSA is semantic structure method that tries to collect text share the same subject into the same topics. Therefore, the number of k topics in LSA is less than and coherence score values are less than LDA. LSA coherence scores obtained using topics modeling algorithms depending on the number of hashtags in the Twitter dataset. Several hashtags have become a limitation in topics Algorithms. In future work, we plan to incorporate other social network features with topic model algorithms to reduce the impact of hashtags numbers.

References
[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” J. Mach. Learn. Res., vol. 3, no. Jan, pp. 993–1022, 2003.
[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” J. Am. Soc. Inf. Sci., vol. 41, no. 6, pp. 391–407, 1990.
[3] T. Hofmann, “Probabilistic latent semantic indexing,” in Proceedings of the 22nd annual ACM SIGIR conference on Research and development in information retrieval, 1999, pp. 50–57.
[4] R. Nugroho, C. Paris, S. Nepal, J. Yang, and W. Zhao, “A survey of recent methods on deriving topics from Twitter: algorithm to evaluation,” Knowl. Inf. Syst., pp. 1–35, 2020.
[5] A. Anantharaman, A. Jadiya, C. T. S. Siri, B. N. V. S. Adikar, and B. Mohan, “Performance evaluation of topic modeling algorithms for text classification,” in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 704–708.
[6] A. Yildirim, S. Usküdarlı, and A. Özgür, “Identifying topics in microblogs using Wikipedia,” PLoS One, vol. 11, no. 3, p. e0151885, 2016.
[7] G. A. Al-Sultany and H. J. Aleqabie, “Enriching Tweets for Topic Modeling via Linking to the Wikipedia,” Int. J. Eng. Technol., vol. 8, no. 1.5, pp. 144–150, 2019.
[8] D. Alvarez-Melis and M. Saveski, “Topic modeling in twitter: Aggregating tweets by conversations,” 2016.
[9] J. Pang, X. Li, H. Xie, and Y. Rao, “SBTM: Topic modeling over short texts,” in International Conference on Database Systems for Advanced Applications, 2016, pp. 43–56.
[10] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, “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, 2016, pp. 165–174.
[11] C. Li, Y. Duan, H. Wang, Z. Zhang, A. Sun, and Z. Ma, “Enhancing topic modeling for short texts with auxiliary word embeddings,” ACM Trans. Inf. Syst., vol. 36, no. 2, pp. 1–30, 2017.
[12] L. Zhu et al., “A joint model of extended LDA and IBTM over streaming Chinese short texts,” Intell. Data Anal., vol. 23, no. 3, pp. 681–699, 2019.
[13] Y. Zuo, J. Zhao, and K. Xu, “Word network topic model: a simple but general solution for short and imbalanced texts,” Knowl. Inf. Syst., vol. 48, no. 2, pp. 379–398, 2016.
[14] A. Steinskog, J. Therkelson, and B. Gambäck, “Twitter topic modeling by tweet aggregation,” in Proceedings of the 21st nordic conference on computational linguistics, 2017, pp. 77–86.
[15] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, “An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit,” Inf. Process. Manag., vol. 57, no. 2, p. 102034, 2020.
[16] W. Cui et al., “Extended search method based on a semantic hashtag graph combining social and conceptual information,” World Wide Web, vol. 22, no. 6, pp. 2589–2610, 2019.
[17] X. Zhu, J. Huang, B. Zhou, A. Li, and Y. Jia, “Real-time personalized twitter search based on semantic expansion and quality model,” Neurocomputing, vol. 254, pp. 13–21, 2017.
[18] M. Prateek and V. Vasudeva, “Improved topic models for social media via community detection using user interaction and content similarity,” in 2016 international FRUCT conference on intelligence, social media and web (ISMW FRUCT), 2016, pp. 1–7.
[19] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, “Exploring topic coherence over many models and many topics,” in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 952–961.