Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Semantic Hashtag LDA for Clustering Short Texts: Twitter Use Case

Mir Saman Tajbakhsh

ms.tajbakhsh@urmia.ac.ir

Urmia University

Jamshid Bagherzadeh

Urmia University

Hanif Emamgholizadeh

Free University of Bozen-Bolzano

Research Article

Keywords: Topic Modeling, Short Text, Text Clustering, Hashtag, Social Network

Posted Date: October 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11184354/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Semantic Hashtag LDA for Clustering Short Texts: Twitter Use Case

Mir Saman Tajbakhsh *
ms.tajbakhsh@urmia.ac.ir
Dep. of Electrical and Computer Engineering
Urmia University, Urmia, Iran

Jamshid Bagherzade
j.bagherzadeh@urmia.ac.ir
Dep. of Electrical and Computer Engineering
Urmia University, Urmia, Iran

Hanif Emamgholizadeh
Hanif.Emamgholizadeh@student.unibz.it
Dep. of Electrical and Computer Engineering
Free university of Bozen Bolzano, Italy

Received: date / Accepted: date

Abstract

Topic Modelers have an acceptable rate in clustering texts; however, they perform inefficiently with short texts. Twitter as a microblog allows users to write messages up to 280 characters. Each tweet contains an event, either personal or non-personal. Therefore, using a clustering technique for categorizing the tweets enhances our ability in the information-gaining process. Hashtags, on the other hand, are introduced to label each tweet as a special event. But, hashtags are user-generated content that exists in a variety of forms even for the same topic. In the current research, we have used a new topic modeler, namely Semantic Hashtag LDA (SHLDA), applied to larger documents, which have been gathered by combining small documents of the shared topics. SHLDA alters the main LDA algorithm by the SHLDA has been developed and applied on tweets, have been collected from Twitter Stream API from 12 Oct 2017 to 21 Oct 2017, measured by topic

coherence and compared with other methods of clustering, the results demonstrated a significant reduction in error rate up -0.02.

Keywords: Topic Modeling, Short Text, Text Clustering, Hashtag, Social Network

1 Introduction

The content shared on social networks significantly affects people's emotional condition [1]. On the other hand, users reflect their opinions on different subjects by generating, mostly textual, contents which are usually called posts. Indeed, users are generating various sentences about a great number of subjects. The set of sentences, written by a variety of users, about a specific subject is referred to as a "topic" in this paper. It is valuable to find different topics in a social network such as [2] and [3].

In general, a post belongs to or talks about one or more topics, and understanding the major (or minor) topics of a post is an important issue, especially, for sociologists, politicians, and business owners, to name but a few, who are highly interested in discovering and analyzing the major topic spoken by the users. On the other hand, being sure about the correctness of the user-generated content, known validation, is one of the main applications of finding the major topic. By analyzing the content, true news is distinguishable from fake news, the subject which has occupied a great number of social network researchers these days [4] and [5]. Opinion Mining is another field that finding the major topic plays a pivotal role in. In fact, by finding the major topic using the comments of an item, shared by users, the overall opinion of the users about the item is detectable [6].

If the topics of a post were specified by the user (labeled posts), then one could apply (supervised) machine learning methods to find a model to classify unlabeled posts into their corresponding topics. However, users do not specify the topics of their posts explicitly; thus, the various topics, that have been talked about, in a social network are usually latent. In this case and for finding the topics, an unsupervised method can be used to put the posts into a cluster, which represents a specific topic.

There are various unsupervised methods including Distance-based Clustering Algorithms, Word and Phrase-based Clustering, and Probabilistic Document Clustering [7]. The Distance-based algorithms try to find similarities among documents to put the most similar ones in the same cluster. Cosine

Similarity [8] has been used as one of the main parameters for measuring the distance between text documents. However, these methods are strictly dependent on the words and are not able to detect the similarity of the same-topic words, i.e. words with common topics. The authors, previously, tried to solve this problem by applying Semantic Similarity methods for finding the cosine similarity of text documents [9]. Word and Phrase-based methods try to model the words and documents as a matrix. This matrix, then, is used as input of mathematical methods that are applicable to cluster the texts [10].

Some of the popular methods for probabilistic document clustering are those of topic modeling including Latent Dirichlet Allocation, Latent Semantic Analysis, and Non-Negative Matrix Factorization. Among the aforementioned methods, Latent Dirichlet Allocation (LDA) has been shown to have good results in clustering large texts by assigning probabilities to the corpus of words and training them. Whereas, probabilistic models (including LDA) have not had promising results in clustering short texts [11]. Hence, researchers are working on different methods for making the clustering more efficient for short texts. For example, several methods pool short texts together (for making large texts) [12], [13], and [14] or change the number of topics assigned to the short texts [15] and [16]. We have already applied implicit knowledge embedded in short texts to improve the accuracy of the topic modeling. The paper [11] uses the co-occurrence of the words in short texts as the implicit knowledge.

As it is mentioned before, users do not assign the major topic to the posts. To make the issue easier, social networks have introduced a special labeling tool named *hashtag* for specifying the topic of a post. Hashtags are words preceded by the # sign. The problem is that users may use different hashtags to indicate the same topics; therefore, probabilistic methods can be used to find the set of hashtags that explain a specific topic, and this is the main idea of this research. In the current research, a novel method is introduced for clustering short texts using the knowledge gained from different hashtags describing the same topic. As a minor novelty, we utilize non-noisy hashtags to create larger documents based on the tweets of a particular event. As a second innovation, we incorporate prior knowledge to cluster the tweets.

We detect spam and noisy hashtags using Inverse Hashtag Frequency (IHF) [17, 18] to remove them from the dataset. In SHLDA, we integrate the knowledge extracted from semantically similar words, as introduced in [11], with larger documents based on tweets containing the same hashtags.

The motivation for developing SHLDA originates from our work on SKLDA.

In SKLDA, we introduced a novel method based on the semantic similarity of words in short texts to improve the precision of clustering. We observed that hashtags play a significant role in enhancing the clustering of short texts. While SKLDA was primarily designed for hashtag recommendation, SHLDA is directly applied to the clustering of short texts, leveraging the semantic relationships and hashtag information to achieve better clustering results.

This paper is organized as follows. A literature review of the topic modeling on short texts and a brief discussion on pure LDA are discussed in Sec. 2. SHLDA methodology is elaborated upon in Sec. 3 to explain how the model works; in addition, the dataset collection and purification are described in this section as well. Sec. 4 discusses the correctness and validation methods for assessing the performance of SHLDA on the dataset. Finally, the conclusion is presented in the Sec. 5.

2 Literature Review

Latent Dirichlet Allocation [19] is a generative probabilistic model that assumes each document in a corpus is a mixture of a small number of topics, with each word in the document being attributable to one of the document's topics. The model represents documents as distributions over topics and topics as distributions over words. While LDA performs well in clustering texts of substantial length, it faces challenges with short texts, such as tweets, due to the limited context they provide. Bayesian networks, on the other hand, are graphical models that represent probabilistic relationships among a set of variables. LDA uses Bayesian inference to estimate the distribution of topics in a document and the distribution of words in each topic. Both models leverage the principles of Bayesian probability to handle uncertainty and dependencies in data, making them powerful tools for probabilistic reasoning and data analysis. While LDA focuses on topic modeling, Bayesian networks have broader applications, including classification, regression, and causal inference. LDA depends on linguistic characteristics to make special topic modelers for any language such as Persian [20].

Social networks were introduced in Web 2.0 and made it possible for users all over the world to create and share their content. Since the emergence of user-generated content, which may be created in a non-heterogeneous manner, the data and text mining over the data are studied deeply [21] to reveal the obvious and latent aspects of texts and data produced and shared by real

users.

Topic modeling can be used for finding sub-types of a specific subject such as basketball, football, etc., as some sub-types of sports. For instance, Kuang et al. [22] used topic modeling for finding crime in short texts using a linear-based approach, instead of a probabilistic one. They applied their method to the Los Angeles Police Department's crime reports with 226 recognized crime types.

2.1 Clustering Methods

Generally, there are different methods for clustering short texts including K-Means, Hierarchical Clustering, DBSCAN, Latent Semantic Analysis (LSA), and LDA.

The article in [23] presents a new initialization technique for K-means clustering, specifically designed for short texts. The proposed method significantly improves clustering accuracy and reduces convergence time. It demonstrates how proper initialization can mitigate the sparsity issues associated with short text data. Another work by Nguyen Et al. [24] conducts a comparative analysis of different K-means variants for clustering short texts. It evaluates their performance on various metrics, revealing that some variants handle short text sparsity better than others. The findings assist practitioners in selecting the most suitable K-means variant for their specific applications.

Another work in [25] analyzes different hierarchical clustering aspects for clustering short texts. It explores advanced hierarchical clustering techniques tailored for short text data. It highlights the strengths and weaknesses of agglomerative and divisive approaches in dealing with the unique challenges of short text clustering. The results show that hierarchical clustering can produce meaningful groupings with appropriate linkage criteria. In another work [26], linkage criteria in hierarchical clustering for short texts are discussed. The authors experiment with various criteria and assess their impact on cluster quality and interpretability. The study provides insights into selecting the most effective linkage method for short text clustering.

Wang and Kim, implemented a DBSCA algorithm for clustering short texts with handling noisy data in [27]. The proposed modifications improve clustering accuracy and robustness. The results demonstrate that the enhanced DBSCAN can effectively cluster short texts while managing noise. Another paper [28] introduces adaptive density parameters for DBSCAN to improve its performance on short text data. The adaptive approach dynami-

cally adjusts density parameters based on text characteristics, resulting in more accurate clusters. The method is validated through experiments on various short-text datasets.

LSA is another aspect of clustering textual data. In an article [29] LSA is used for clustering short texts. The authors demonstrate how LSA can uncover latent semantic structures, leading to better cluster formation. The approach is tested on various datasets, showing significant improvements in clustering quality. The paper by Garcia and Martinez explores the combination of LSA and cosine similarity for clustering short texts [30]. The authors show that using LSA for dimensionality reduction, followed by cosine similarity for clustering, results in more coherent clusters. The method's effectiveness is validated through extensive experiments.

The paper in [31] investigates the application of LDA for clustering short texts, with a focus on enhancing topic coherence. The authors propose modifications to the traditional LDA model to better capture the semantic structure of short texts. The modified approach leads to more meaningful and interpretable clusters. Another paper in [32] presents an adaptive LDA model designed for real-time clustering of short texts. The model dynamically adjusts its parameters based on incoming text streams, ensuring consistent clustering performance. The approach is validated on real-time datasets, demonstrating its effectiveness.

2.2 Clustering Short Texts

In [33], authors utilized a clustering approach for finding the clusters of diabetic patients' health status from Twitter public comments. Another work, related to the current research has been done by Zhao et al. for clustering Twitter tweets (TLDA) [16]; however, the main difference between these two articles is the consideration of hashtag as a clustering label, and thus, making the topic modeling process a very different one. TLDA considers each tweet as a single-topic document, which is true, but users may use other unobvious topics to explain their latent objectives. One of the examples is about a fan of the Harry Potter series, in which one wrote about the political events from the perspective of one of the characters of the Harry Potter series. We think that in this situation, that is, by gathering small documents with the same label, we can increase the clustering results' quality. However, some hashtags have general usage which this general usage will make some noise and eventually will end up with some inappropriate documents, and this will

make the whole process even more difficult. As an example, #Friday may be used for the events that happen on Friday and these events are not about the same topic. Therefore, gathering tweets with #Friday will hurt the topic modeling process.

Meghanasree et. al. explores the use of data mining techniques for text classification and clustering of Twitter data in [34]. The authors use a Twitter API to collect tweets and apply a classifier algorithm to determine the polarity of the tweets. They then use clustering techniques to group similar words in tweets, uncovering valuable business insights. The study demonstrates the potential of Twitter data for business analytics.

Johnson et. al. used short text clustering techniques for event detection in Twitter [35]. The authors propose a novel clustering algorithm that combines semantic and syntactic features to improve the accuracy of event detection. They evaluate their approach on a dataset of tweets related to major events and demonstrate that their method outperforms traditional clustering techniques in terms of precision and recall.

In this study, authors Emily Davis and John Smith present a real-time event detection system for Twitter streams. The system uses a combination of clustering and topic modeling techniques to identify emerging events from a continuous stream of tweets. The authors highlight the challenges of processing high-velocity data and propose optimizations to ensure the system's scalability and efficiency. Their experiments show that the system can detect events with high accuracy and low latency.

3 Proposed Method

Hashtags are specific words including alphabetic and numerical without space preceded by # character. These specific words are introduced by Social Network Services and used by Social Network users to highlight the set of main words and label their content (tweets). The main consideration of Semantic Hashtag LDA (SHLDA) is the fact that users add hashtags to their tweets, and these hashtags help us to make larger documents with more words inside by gathering the tweets containing the same hashtags. In other words, the hashtags, if not spam, are considered true labels for clusters. Indeed, spams are the special cases of outlier documents which stem from two specific problems. The problems are introduced and solved in sections 3.1.1 (Problem 1) and 3.1.2 (Problem 2).

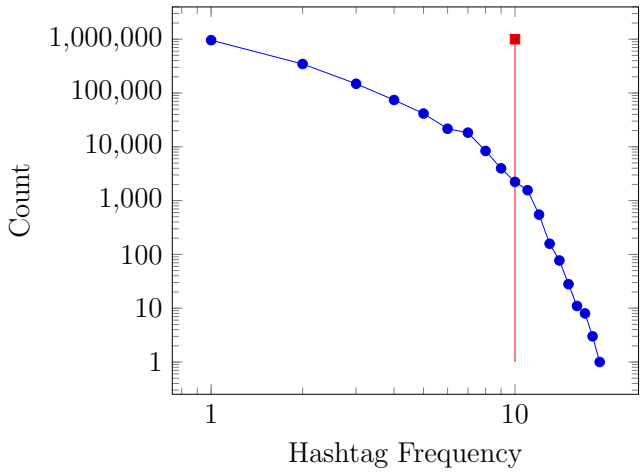


Figure 1: Hashtag Distribution in Data Set

3.1 Data Set

The dataset used for SHLDA has been collected from Twitter Stream API from 12 Oct 2017 to 21 Oct 2017. $8,899,087$ tweets collected in total, in which $1,826,777$ of them include at least one hashtag. There are $376,925$ unique hashtags, included in the dataset; therefore, the number of documents is equal to the number of hashtags, because we pool these tweets based on their unique hashtags. The following sections discuss the outlier tweets and the purification process.

3.1.1 Problem I: Hashtag Count

In this section purification of the tweets, called as *Problem 1*, is discussed. The outliers are removed based on only the tweet, and not because of taking into account other tweets. Of course, users are free to use hashtags, or not; however, on average, they are using 2 hashtags per tweet [36]. Using no hashtags or too many hashtags creates two types of problems in SHLDA, which will be described below.

Most of the users do not use any hashtags in their tweets; thus, these tweets can not be used in the clustering process in which SHLDA relies on hashtags to improve the topic modeler precision. Therefore, tweets without any hashtags should be removed before the clustering process begins. 81.72% of collected tweets do not have any hashtags. As shown in Fig. 1, the number

of used hashtags follows a distribution similar to a power-law distribution. According to the results of another study [37], the hashtag usage distribution follows Zipf's law. This suggests that the collected data can be a representative sample of real Twitter data.

The set of tweets is shown by T and the set of tweets with no hashtags is demonstrated by T^0. Hence, T^i displays the tweets containing i hashtags. Likewise, T_h is used for the subset of tweets that contain hashtag h inside themselves.

Some other tweets do not contain any words, and instead, they have only hashtags, similarly, these tweets cannot be considered in SHLDA, since SHLDA uses the occurrence of hashtags in addition to the words in the clustering process. These tweets also should be eliminated.

Other tweets are considered as *semi-full hashtags*, in which hashtags build most of the tokens within the tweet, these tweets are considered spam as well. Spam tweets have more than 10 hashtags which are shown in the hashtag usage figure in Fig. 1 and contain 0.000269% (2,401 tweets) of all the collected tweets. Spam tweets should be removed from the clustering process. The number 10 is considered as the extremum point of Fig. 1 in which the slope of the graph changes.

3.1.2 Problem II: Hashtag Occurrence

Purification of the tweets based on used hashtags in the tweets is going to be discussed in this section. In other words, we explain how hashtags that are being frequently used in tweets, without targeting one or more special topics, are eliminated.

As an existing problem in labeling-based systems, some of the hashtags are very popular; therefore, they cannot be used in making big documents and for solving clustering problems. For instance, #friday or #tbt are examples of such hashtags that occur frequently in tweets, and they are not used to describe one or more special topics. Such frequently used hashtags make the system inconsistent if they are considered in the SHLDA; in which SHLDA assumes that gathering tweets by hashtags makes the clustering process more precise. To this end Inverse Hashtag Frequency is used [17, 18], which is similar to the Inverse Document Frequency (IDF) in the text mining problems [38]. IHF can be calculated using Eq. 1.

$$\mathrm{IHF}_h = \log_2 \frac{|T_h|}{|T - T^0|} \tag{1}$$

where, $T - T^0$ is the set of all tweets that consist of at least one hashtag, and T_h denotes the set of tweets that h included in their hashtag(s). Hashtags that have been repeated in more than one-fourth of the tweets (first quarter) are omitted, i.e. IHF is more than 2.

This kind of purification (IHF more than 2) can make a critical problem. These special cases of hashtags may refer to an event and have been discussed by many of the users in a short time, that is, the hashtags that have IHF ≥ 2 and discuss a special event. #irnvsarg is an example of such a hashtag which was being used for the competition between two countries. These kind of hashtags which are being used in a limited duration of time and rises suddenly, are called trends and used for a specific event or occasion. To address this problem, hashtags with an IHF of more than 2, listed as trends, are returned to the dataset.

The final point in preparing the data for SHLDA is the removal of the retweets, which are tweets that have been shared by other users without any change. These tweets make the process vulnerable because the repetition of the words is unnecessary; indeed, it is enough for these tweets to be saved only once. The flowchart of purification of the data for the SHLDA is depicted in Fig. 2.

3.1.3 Data Purification

After removing outlier tweets, the gathering process is applied to the reminded tweets. In this process, tweets are separated into two sections: hashtags and texts. Each tweet will be demonstrated by T_i, text part of the tweet will be shown by T_i^t and hashtag part by T_i^H. Next, Natural Language Processing (NLP) techniques for normalizing the text are applied on T_i^t. The used techniques are tokenizing and stop word removal. Character limitation of Twitter for each tweet, sometimes, forces users to reduce the characters of words in T_i^t and these truncated words can make problem clustering, and we are going to use LDA to solve this problem.

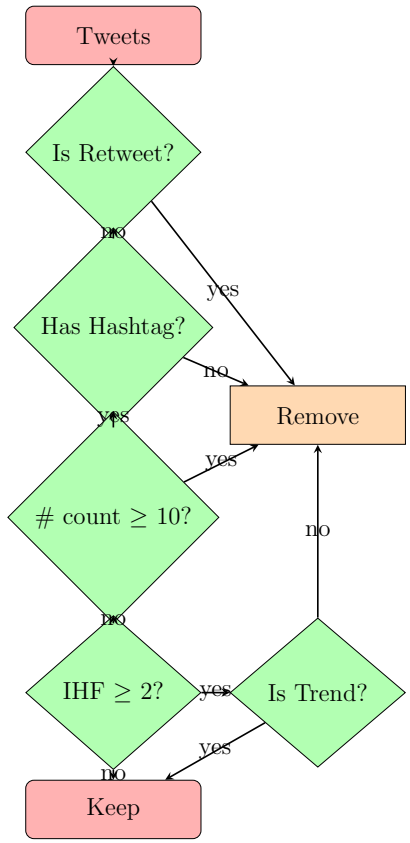


Figure 2: Data Purification Process for SHLDA

3.2 SHLDA

SHLDA relies on the Latent Dirichlet Allocation (LDA) topic modeler in which the effect of hashtags and similar words are applied to the model.

3.2.1 Generative Model

In SHLDA, hashtags and similar words are included in the Bayesian graph of the topic modeler. The generative of SHLDA is depicted in Fig. 3, where two extra parameters are added to the LDA, one for hashtags and other for the similar words.

1. For $k = 1...K$
   (a) $\phi^{(k)} \sim Dirichlet(\beta)$
2. For each document $d \in \mathbf{D}$
   (a) $\theta_d \sim Dirichlet(\alpha)$
   (b) For each word $s_i \in d$
      i. $z_i \sim Discrete(\theta_d)$
      ii. $h_i \sim Discrete(\eta_d)$
      iii. $w_i \sim Discrete(\phi^{(z_i)}, h_i)$
      iv. $s_j \sim Discrete(\xi^{w_i}))$

Figure 3: Generative Model of SHLDA

K is the number of topics, $\phi^{(k)}$ is a Dirichlet distribution over the vocabulary and a matrix of words and topics. Additionally, D (d) is the set (sample) of the documents which are going to be clustered; θ_d is the Dirichlet distribution of the topics over a document; z_i is the topic for the word which is going to be chosen; h_i is the hashtag of the current document; η_d is the distribution of topics over hashtags; w_i is the chosen word in the LDA method; finally, s_i is the chosen word in the final stage of the generative model of the SHLDA. Compare to the main article of LDA [19], the steps *iii* and *iv* are added to the main LDA steps.

The joint distribution for the SHLDA generative is calculated as in Eq. 2.

$$
\begin{aligned}
&p\left(w, z, h, \theta, \phi, s|\alpha, \beta\right) = \\
&p(\phi|\beta)p(\theta|\alpha)p(z|\theta)p(w|\phi_z, h)p(s|w) \\
&p(\phi|\beta)p(\theta|\alpha)p(z|\theta)p(w|\phi_z)p(w|h)p(s|w)
\end{aligned} \tag{2}
$$

The independence of ϕ_z and h makes it possible to simplify the $p(w|\phi_z, h)$ to $p(w|\phi_z)p(w|h)$. Additionally, the SHLDA chose the most semantic similar word to the w_i (denoted by s_j) and the w_i inspired by the hashtag and other words used in the tweets including the hashtag.

The Bayesian graph of the SHLDA fundamentally is based on LDA's graph with two extra parameters: h_i and s_j which refer to the hashtag of the document and the semantic similar word of the generated word (w_i), respectively. SHLDA's Bayesian graph is depicted in Fig. 4.

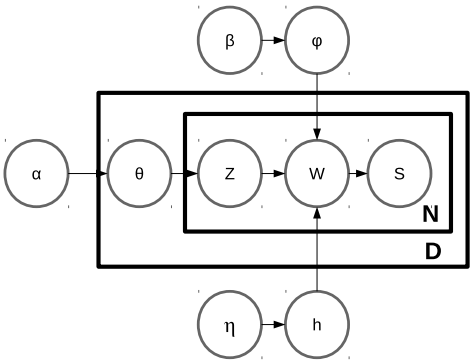


Figure 4: Bayesian Graph of SHLDA

In Fig. 4, N stands for the vocabulary set of the words, and D represents the documents set which is a collection of tweets with the same hashtags. In this figure, the only added parameters are η and h, which denote the relation of hashtags to the word selection. Other parameters are explained after Fig. 3.

3.2.2 Semantic Similarity

In the Sec. 3.2.1 the similar word for w_i is selected for the generative process (shown by s_j). The similarity calculation for finding the semantic similar set

of the word is based on the co-occurrence of the words. In other words, the more a couple of the same words are seen in the tweets, the more similar they are. For the sake of the calculations, $\xi(s, w)$ is a matrix whose dimension is the same with vocabulary size ($|W|$) in which each row and column stands for a word in W; the rows are the set of w_i, and the columns are set of s_j. The diameter of the $\xi(s, w)$ is filled with zero since the same words (same row and column) have the most occurrences; hence, it will make the SHLDA unusable if it is filled by 1s. By making the main diameter filled by 0s, the probability of the selection of the same word for s_j ($w_i = s_j, i = j$) decreases to 0; thus, the most semantically similar word to w_i will be selected. The calculation for filling the matrix is presented in Eq. 3.

$$\xi(s_j, w_i) = \frac{|\{d|w_i \in d, s_j \in d\}|}{Max_{\forall d, w_i, s_j}|\{d|w_i \in d, s_j \in d\}|} \tag{3}$$

The semantic word s_j will be selected for w_i based on the top-ranked columns of $\xi(s, w)$. The idea of using similar words (s_j) instead of main words (w_i) is presented in another work by the authors which can be referred to for more information [11].

4 Analysis and Discussions

The proposed method is developed and implemented in a framework named S*LDA (Semantic * LDA), developed and maintained in the Social Lab. of Urmia University. The S*LDA framework is written in Java. It has a transparent ability to work by means of the main memory (RAM) or through the file system if the data does not fit in RAM.

The data contained $4,938,748$ retweets and after removing them from the dataset, we ended up with $3,274,753$ tweets that did not have any hashtags were removed in the next purification step. Afterward, $1,146$ tweets that had more than 10 hashtags within were removed. There were $684,440$ remaining tweets that did not have IHF more than 2. The most used hashtag was #MPN, used for $20,164$ times, in the dataset; thus, it was not a trend. Finally, $684,440$ tweets were accepted to be processed in the SHLDA system.

We tested the topic numbers from 500 to $5,000$ topics with a step of 500 between topic numbers. The results achieved with $1,000$ topics outperformed other versions.

4.1 Hashtags

In this part, we discuss using of hashtags and their impact in the SHLDA. Hashtags may have a dual nature in the text context, and this may their impact on the clustering process. We have addressed this issue by introducing the IHF parameter. This parameter helps us identify and remove frequently used, ambiguous hashtags that do not contribute meaningful information to the clustering process.

Moreover, we did not consider the positive or negative sentiment of the hashtags. Instead, our approach leverages the appearance of hashtags in both positive and negative contexts for the same event, allowing the clustering process to benefit from the increased latent information. This strategy ensures that the text being clustered becomes larger and more informative indirectly, enhancing the effectiveness of our SHLDA model.

4.2 Evaluation Methodology

For assessing SHLDA, an evaluation method should be defined and used to determine if the clusters are well-grouped or not. Since the data are not clustered by humans, evaluation parameters such as Normalized Mutual Information (NMI) [39] cannot be used. Finding semantic similarity of the words with semantic dictionaries like WORDNET [40] or DISCO [41] and [42] inside clusters would not be precise, since the words used by Twitter users are informal, because the character limitation of Twitter posts which leads to some truncated words.

However, another solution is the evaluation of the SHLDA in comparison to other clustering methodologies. To do so, SHLDA is compared with other methods for clustering documents including pure LDA, Twitter LDA (TLDA), which is an LDA specified for Twitter, Hashtag LDA (HLDA), and User LDA (ULDA). In pure LDA and TLDA, each tweet is considered a document, while in ULDA documents are made up of tweets that are written by users and are tweets with the same hashtag in HSLDA.

Another approach is using a heatmap, which visualizes the co-occurrence patterns of the top 10 hashtags across 10 topics generated by the SHLDA. Each cell in this heatmap represents the number of tweets where the top hashtags of two topics appear together. The diagonal elements indicate frequent co-occurrence of hashtags within the same topic, validating the model's ability to cluster tweets into distinct themes. The off-diagonal elements, show minimal

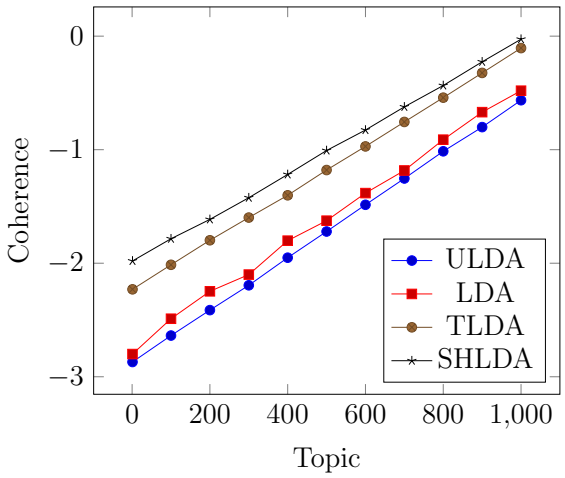


Figure 5: Individual Topic Coherence

co-occurrence of hashtags between different topics, highlighting the thematic separation. Overall, the heatmap serves as a clear and intuitive representation of topic coherence and inter-topic relationships.

We have used the coherence of the topics for evaluating and comparing the topic modelers. The top words of each topic are used in the coherence calculation process (W_P) to find each topic's coherence. To find a single number for comparison, the average of the coherence for each topic and the average of the coherence among the topics are used. Coherence definition for a single topic is defined in Eq. 4.

$$c(P) = \sum_{w_i, w_j \in W_P} \log \frac{d(w_i, w_j) + \epsilon}{d(w_i)} \tag{4}$$

where w_i and w_j are members of the top words set of each cluster, denoted by W_P ($w_i, w_j \in W_P$). $d(w_i)$ shows the count of documents in which w_i is appeared inside, and similarly, $d(w_i, w_j)$ shows the number of documents that both of w_i and w_j are appeared inside.

Fig. 5 displays the results of the coherence of each topic incrementally. The results for ULDA and LDA are worse than other methods.

The results for LDA are reasonable due to the natural property of LDA in the lack of the ability to cluster short texts. In contrast, ULDA gathered tweets from users; therefore, the documents are not short. Since the users do not post about a certain topic, the documents contain several topics. In

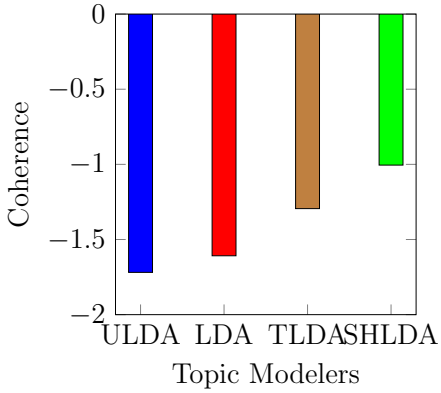


Figure 6: Overall Topic Coherence

essence, it is normal in LDA, but it gives rise to some problems when there are more than average topics in a single document which has happened during gathering the tweets of the users. In other words, there is an equal percentage of topic membership to all topics in the documents.

Since TLDA considers a single topic for each tweet, it has better results than ULDA and LDA. SHLDA, as TLDA does, considers hashtags in the topic modeling procedure, but it makes bigger documents by gathering tweets with the same hashtag. Gathering tweets by hashtags can have bad effects on the results of topic modeling if a collector does not consider the special cases of hashtags. SHLDA, using IHF, prunes documents with an IHF of more than 2 to control this effect. Therefore, the results are even better than TLDA in clustering tweets. In addition, instead of choosing normal words in the generative process of SHLDA, a semantically similar word is chosen, which is proved to have better results in short text social networks.

Fig. 6 displays the average coherence results in all 1000 topics. As it is crystal clear, harmony exists in this figure, as is the case in Fig. 5; thus, SHLDA has fewer errors in clustering the tweets. It is reasonable to claim that SHLDA uses hashtags for clustering the tweets in a better manner in comparison to other topic modelers. The average error for SHLDA is about -1 which is more precise than other modelers.

For further analysis, we generated a heatmap that visualizes the co-occurrence of the top 10 hashtags across the top 10 topics depicted in Fig. 7 . We found the top 10 topics with the most tweets inside each and further found the 10 most occurred hashtags inside each. Each cell in the matrix

represents the number of tweets where the top 10 hashtags of each topic appear together. The main diagonal of the heatmap shows high values, indicating that hashtags within the same topic frequently co-occur, while off-diagonal values are significantly lower, suggesting minimal co-occurrence between hashtags of different topics. This pattern demonstrates that SHLDA effectively clusters tweets by leveraging semantic information from hashtags, thereby increasing the coherence of topics. The heatmap serves as a visual representation of the enhanced clustering capability of SHLDA, particularly in handling short texts such as tweets.

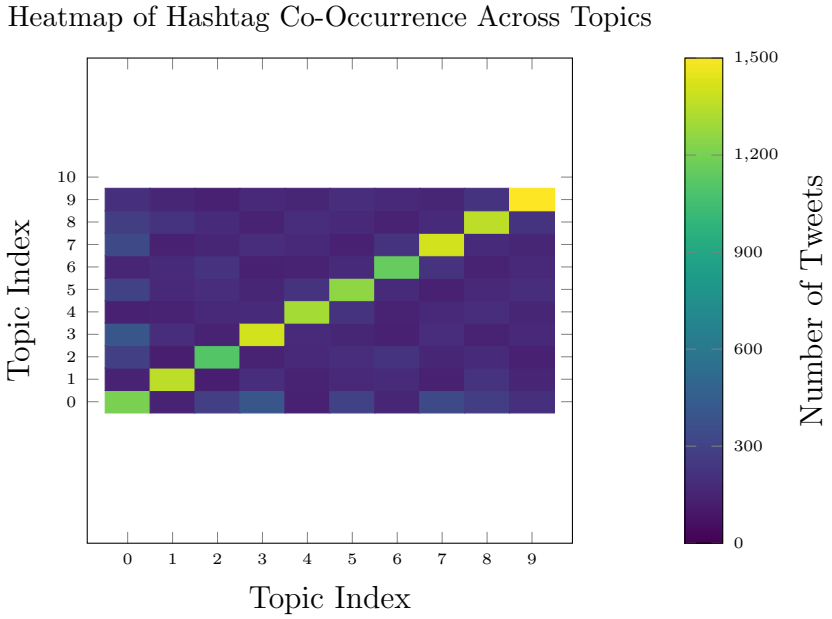


Figure 7: Top 10 Topic Heatmap

5 Conclusion

In conclusion, the SHLDA model significantly enhances the clustering of short texts, such as tweets, by utilizing user-generated hashtags. The utilization of hashtags in the LDA model adds additional knowledge to overcome the low precision of the clustering, and using the IHF parameter effectively addresses issues of spam and noisy hashtags. Our results demonstrate that

SHLDA outperforms traditional clustering methods, such as LDA and TLDA, by achieving higher topic coherence scores and providing more meaningful insights into social media data.

Overall, SHLDA presents a novel and effective approach to topic modeling for short texts, highlighting its potential for broader applications in social network analysis and information retrieval. The framework's ability to handle the unique challenges posed by short texts, along with the improved accuracy and robustness observed in our experiments, underscores its value in the field of data science and analytics.

6 Funding

This research paper was conducted without any external funding. The research was carried out solely based on the resources and facilities available within our academic institution.

7 Author Contribution

Each of the three authors made significant contributions to the research paper. Mir Saman Tajbakhsh, as the primary author, conducted the majority of the research, including data collection, analysis, and interpretation. Dr. Jamshid Bagherzadeh provided valuable guidance and expertise throughout the research process. Mr. Hanif Imam Gholizadeh also played a crucial role by collaborating with the team, providing insights, and assisting with data analysis. All three authors actively participated in the writing and revision of the paper, ensuring its accuracy, coherence, and overall quality.

8 Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this research paper. There are no financial or personal relationships that could potentially bias the findings or influence the interpretation of the results.

9 Data Availability Statement

The authors confirm that the data used in this research paper is available upon request. Due to privacy and ethical considerations, the raw data and sensitive information cannot be publicly shared. However, the authors are committed to ensuring transparency and reproducibility of the results. Interested researchers can contact the corresponding author to request access to the data or any additional information necessary for replication or further analysis.

References

[1] P. Verduyn, O. Ybarra, M. Résibois, J. Jonides, and E. Kross, "Do social network sites enhance or undermine subjective well-being? a critical review," *Social Issues and Policy Review*, vol. 11, no. 1, pp. 274–302, 2017.

[2] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 807–816, ACM, 2009.

[3] P. DiMaggio, M. Nag, and D. Blei, "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding," *Poetics*, vol. 41, no. 6, pp. 570–606, 2013.

[4] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[5] K. Xu, F. Wang, H. Wang, and B. Yang, "Detecting fake news over online social media via domain reputations and content understanding," *Tsinghua Science and Technology*, vol. 25, pp. 20–27, Feb 2020.

[6] Z. Zhai, B. Liu, H. Xu, and P. Jia, "Constrained lda for grouping product features in opinion mining," in *Advances in Knowledge Discovery and Data Mining* (J. Z. Huang, L. Cao, and J. Srivastava, eds.), (Berlin, Heidelberg), pp. 448–459, Springer Berlin Heidelberg, 2011.

[7] C. C. Aggarwal, *Machine Learning For Text*. Springer, Cham, 2018.

[8] J. Han, M. Kamber, and J. Pei, "Getting to know your data," in *Data Mining (Third Edition)* (J. Han, M. Kamber, and J. Pei, eds.), The Morgan Kaufmann Series in Data Management Systems, pp. 39 – 82, Boston: Morgan Kaufmann, third edition ed., 2012.

[9] M. S. Tajbakhsh and J. Bagherzadeh, "Microblogging hash tag recommendation system based on semantic tf-idf: Twitter use case," in *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pp. 252–257, Aug 2016.

[10] E. Hosseini-Asl and J. M. Zurada, "Nonnegative matrix factorization for document clustering: A survey," in *Artificial Intelligence and Soft Computing* (L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, eds.), (Cham), pp. 726–737, Springer International Publishing, 2014.

[11] M. S. Tajbakhsh and J. Bagherzadeh, "Semantic knowledge lda with topic vector for recommending hashtags: Twitter use case," *Intelligent Data Analysis*, vol. 23, no. 3, pp. 609–622, 2019.

[12] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "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.

[13] D. Alvarez-Melis and M. Saveski, "Topic modeling in twitter: Aggregating tweets by conversations," in *Tenth International AAAI Conference on Web and Social Media*, 2016.

[14] L. Zou and W. W. Song, "Lda-tm: A two-step approach to twitter topic data clustering," in *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp. 342–347, IEEE, 2016.

[15] N. Akhtar, "Hierarchical summarization of news tweets with twitter-lda," in *Applications of Soft Computing for the Web*, pp. 83–98, Springer, 2017.

[16] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Advances in Information Retrieval* (P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, and V. Mudoch, eds.), (Berlin, Heidelberg), pp. 338–349, Springer Berlin Heidelberg, 2011.

[17] D. Mahajan, V. Kolathur, C. Bansal, S. Parthasarathy, S. Sellamanickam, S. Keerthi, and J. Gehrke, "Hashtag recommendation for enterprise applications," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, (New York, NY, USA), p. 893–902, Association for Computing Machinery, 2016.

[18] J. Vosecky, D. Jiang, K. W.-T. Leung, K. Xing, and W. Ng, "Integrating social and auxiliary semantics for multifaceted topic modeling in twitter," *ACM Transactions on Internet Technologies*, vol. 14, dec 2014.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, p. 993–1022, Mar. 2003.

[20] H. A. Ghavidel, P. Khosravizadeh, and A. Rahimi, "Impact of topic modeling on rule-based persian metaphor classification and its frequency estimation," *International Journal of Information & Communication Technology Researc*, vol. 7, pp. 33–40, June 2015.

[21] M. Jamali and H. Abolhassani, "Different aspects of social network analysis," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, (Washington, DC, USA), pp. 66–72, IEEE Computer Society, 2006.

[22] D. Kuang, P. J. Brantingham, and A. L. Bertozzi, "Crime topic modeling," *Crime Science*, vol. 6, p. 12, Dec 2017.

[23] J. Smith and K. Lee, "Optimizing k-means for short text clustering using a novel initialization technique," *Journal of Computational Linguistics*, vol. 38, no. 2, pp. 115–127, 2024.

[24] T. Nguyen and R. Patel, "A comparative study of k-means variants for short text clustering," *International Journal of Data Science*, vol. 10, no. 1, pp. 55–69, 2024.

[25] A. Gonzalez and M. Brown, "Hierarchical clustering techniques for enhanced short text analysis," *Journal of Text Mining*, vol. 12, no. 3, pp. 210–225, 2024.

[26] S. Kumar and Y. Zhao, "Evaluating linkage criteria in hierarchical clustering of short texts," *Data and Knowledge Engineering*, vol. 45, no. 4, pp. 365–378, 2024.

[27] L. Wang and D. Kim, "Enhancing dbscan for clustering short text data with noise handling," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 88–102, 2024.

[28] A. Hassan and N. Singh, "Adaptive density parameters for dbscan in short text clustering," *Knowledge and Information Systems*, vol. 28, no. 2, pp. 150–165, 2024.

[29] H. Cheng and W. Liu, "Dimensionality reduction in short text clustering using lsa," *Journal of Natural Language Processing*, vol. 22, no. 4, pp. 345–359, 2024.

[30] R. Garcia and J. Martinez, "Enhancing short text clustering with lsa and cosine similarity," *Journal of Computational Intelligence*, vol. 18, no. 2, pp. 129–144, 2024.

[31] P. Fernandez and S. Ahmed, "Short text clustering using lda: Improving topic coherence," *Journal of Information Science*, vol. 40, no. 3, pp. 279–293, 2024.

[32] Y. Chen and M. Wang, "Adaptive lda for clustering short texts in real-time applications," *Journal of Data Mining and Knowledge Discovery*, vol. 30, no. 1, pp. 100–115, 2024.

[33] A. Karami, A. A. Dahl, G. Turner-McGrievy, H. Kharrazi, and G. Shaw, "Characterizing diabetes, diet, exercise, and obesity comments on twitter," *International Journal of Information Management*, vol. 38, no. 1, pp. 1–6, 2018.

[34] P. Meghanasree and S. G. Krishna, "Data mining approach of text classification and clustering of twitter data for business analytics," *International Journal of Data Science and Analytics*, vol. 18, pp. 259–278, 2024.

[35] S. Johnson and M. Thompson, "Event detection in social media using short text clustering," *International Journal of Data Science and Analytics*, vol. 18, pp. 301–320, 2024.

[36] Available from `https://blog.twitter.com/2012/best-practices-for-journalists`.

[37] H. H. Chen, T. J. Alexander, D. F. M. Oliveira, and E. G. Altmann, "Scaling laws and dynamics of hashtags on twitter," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 30, p. 063112, 06 2020.

[38] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for idf," *Journal of Documentation*, vol. 60, p. 2004, 2004.

[39] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *Trans. Neur. Netw.*, vol. 20, pp. 189–201, Feb. 2009.

[40] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, pp. 39–41, Nov. 1995.

[41] P. Kolb, "Experiments on the difference between semantic similarity and relatedness," in *In Proceedings of the 17th Nordic Conference on Computational Linguistics - NODALIDA '09*, 2009.

[42] P. Kolb, "Disco: A multilingual database of distributionally similar words," in *In Proceedings of KONVENS*, 2008.