RESEARCH

A Clustering-based Topic Model using Word Networks and Word Embeddings

Wenchuan Mu[1], Kwan Hui Lim[2*], Junhua Liu[2,3], Shanika Karunasekera[4], Lucia Falzon[4] and Aaron Harwood[4]

[*]Correspondence: kwanhui_lim@sutd.edu.sg
[2]Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore
Full list of author information is available at the end of the article

Abstract

Online social networking services like Twitter are frequently used for discussions on numerous topics of interest, which range from mainstream and popular topics (*e.g.*, music and movies) to niche and specialized topics (*e.g.*, politics). Due to the popularity of such services, it is a challenging task to automatically model and determine the numerous discussion topics given the large amount of tweets. Adding on this complexity is the need to identify these topics with the absence of prior knowledge about both the types and number of topics, while having the requirement of the relevant technical expertise to tune the numerous parameters for the various models. To address this challenge, we develop the Clustering-based Topic Modelling (ClusTop) algorithm that first constructs different types of word networks based on different types of n-grams co-occurrence and word embedding distances. Using these word networks, ClusTop is then able to automatically determine the discussion topics using community detection approaches. In contrast to traditional topic models, ClusTop does not require the tuning or setting of numerous parameters and instead uses community detection approaches to automatically determine the appropriate number of topics. The ClusTop algorithm is also able to capture the syntactic meaning in tweets via the use of bigrams, trigrams, other word combinations and word embedding techniques in constructing the word network graph, and utilizes edge weights based on word embedding. Using three Twitter datasets with labelled crises and events as topics, we show that ClusTop outperforms various traditional baselines in terms of topic coherence, pointwise mutual information, precision, recall and F-score.

Keywords: Topic Modelling; Clustering; Word Embedding; Twitter; Microblogs; Social Networks

1 Introduction

Twitter is a popular microblogging service that is prevalent and widely used in everyday life, with a high volume of 500 million tweets posted on a daily basis [1]. On microblogging services, such as Twitter, users frequently perform discussions and debates on topics of interest, ranging from mainstream and popular topics (*e.g.*, movies, TV, music, entertainment) to niche and specialized topics (*e.g.*, politics, religion, current affairs). The capability to detect and understand the discussions about these topics are useful for numerous purposes, such as understanding the general sentiments and trends of these topics, and recommending accurate and relevant content. However, the large volume and high posting frequency of tweets makes it a significant challenge for users to effectively understand the discussion topics in these tweets [2, 3].

A popular approach is to utilise topic modelling algorithms to automatically detect the topics discussed in a set of traditional text-based documents, such as news articles, academic papers, etc. In such algorithms, the output is a set of keywords denoting the topics that are relevant to each document. Examples of topic modelling algorithms are the original Latent Semantic Analysis [4], Probabilistic Latent Semantic Analysis [5] and latent Dirichlet allocation [6]. These algorithms were developed mainly for topic modelling on traditional and large documents such as news articles or papers [7, 8]. The advent of microblogging services has led to the widespread use of short documents (*i.e.*, tweets) in social media, which traditional topic modelling algorithms do not work well on. In response, various researchers have proposed variants of these traditional topic models, based on various types of aggregation schemes to combine a set of tweets as larger documents [9, 10]. While latent Dirichlet allocation and its variant have been shown to model topics well for traditional documents, the number of topics needs to be defined in advance and they do not account for the syntactic structure of sentences.

In this work, we aim to overcome these limitations by introducing a topic modelling algorithm that is able to automatically determine the appropriate number of topics. Our proposed algorithm is based on the adaptation of community detection algorithms on a network graph where vertices are words and edges are relations between words. Our algorithm is also able to capture the syntactic nature of language via the use of bigrams, trigrams, other word combinations and different word embeddings in constructing our word network graph. In addition, we perform an empirical study to better examine the different types of network graphs based on the types of nodes, edges and embedding techniques, and its effect in terms of the accuracy and quality of topics detected.

1.1 Main Contributions

In this paper, our main contributions are as follows:[1]

1 We propose the Clustering-based Topic Modelling (ClusTop) algorithm that makes use of community detection approaches for modelling topics on Twitter using a word network graph. In this word network graph, nodes represent different definitions of words and phrases and edges represent either word/phrase co-occurrences or the similarity distances between words based on embeddings. Unlike more traditional topic models, ClusTop automatically determines the number of topics by maximizing a modularity score among words in the network.

2 In addition to using a traditional co-word usage network, we experiment with different variants of our ClusTop algorithm based on numerous definitions

[1] This paper is an extended version of [11], with the addition of more than 40% new materials. These additional materials include: (i) an updated literature review to include more recent works; (ii) a more detailed description of our proposed approach; (iii) a new algorithm that utilizes various word embeddings and distance measures between words; (iv) additional experiments and evaluations; (v) a more in-depth discussion of the results and our main findings.

of words (unigrams, bigrams, trigrams, hashtags, nouns from part-of-speech tagging), types of relations (word co-occurrence frequency and word embedding similarity distance) and different aggregation schemes (individual tweets, hashtags and mentions). In addition, we also propose variants based on different word embeddings techniques where edges are weighted based on the similarity distances between different words.

3 Using three Twitter datasets with labelled topics, we evaluate ClusTop and its variants against various LDA baselines based on measures of topic coherence, pointwise mutual information, precision, recall and F-score. Experimental results show that ClusTop offers superior performance based on these evaluation metrics, compared to the various baselines.

1.2 Structure and Organization

The rest of this paper is structured as follows. Section 2 discusses key literature on studying topics on microblogs and topic modelling algorithms. Section 3 describes our ClusTop algorithm. Section 4 outlines our experimental methodology in terms of the dataset used, baseline algorithms and evaluation metrics. Section 5 highlights the results from our evaluation and discusses our main findings. Section 6 concludes this paper and highlights possible future directions for this work.

2 Related Work

In this section, we discuss two main areas of research related to our work, namely the study of topics on microblogs and general topic modelling algorithms.

2.1 Studying Topics using Communities

The most closely related works to our proposed approach are those that make use of community detection techniques for understanding and studying topics on microblogs. As such, we discuss main works that utilise such techniques. Towards the effort to better understand research themes in the Human Computer Interaction domain, Liu et al. [12] used hierarchical clustering on co-keywords usage in academic papers to identify the main research clusters across two different time periods. Researchers have also proposed approaches for identifying communities that frequently interact about common interest topics using various types of community detection algorithms. These approaches are based on topological links such as friendship networks among users and celebrities [13] and interaction links in the form of explicit mentions of other users [14]. Researchers like [15] and [16] have also used community detection algorithms on word networks to identify topics with a focus on network analysis and visualization, and detection of spammer topics, respectively. Fried et al. [17] used topic modelling on a series of food-related tweets to understand health information such as overweight rate and diabetes rate. Others like Surian et al. [18] and [19] combined the use of topic modelling algorithms with community detection algorithms to characterize discussions relating to vaccines on Twitter, study discussion topics of Italian users, respectively.

2.2 General Topic Modelling Algorithms

Also relevant to our work are those that proposed various types of topic modelling algorithms, of which latent Dirichlet allocation (LDA) is a particularly popular one with many variants being proposed. As such, we next discuss a series of works that utilizes the popular LDA for proposing new variants that are targetted for use on microblogs and other forms of short text, as well as the application of LDA on various types of social media. LDA [6] is a popular topic model that is used to determine the set of latent topics associated with a set of documents. Each document is usually represented as a bag-of-words in LDA, with each topic modelled by a distribution of words, and each document is assigned a distribution of topics via a generative process. Variants of bag-of-words, such as keeping nouns only or removing stop words, improve topics' semantic coherence [20, 21]. LDA is sometimes accompanied by other representation structures. Structural relationships among social texts in a discussion tree have been added to LDA as context information to alleviate data sparsity and noise [22]. When word co-occurrences are lacking, distributional word embedding captures semantic and syntactic correlations among words [23]. It helps discover interpretable topics even with large vocabularies that include rare words and stop words [24, 25]. Domain-specific semantic relationships of words are useful in areas such as clinical predictive modelling [26] and restricting keywords to specific predefined topics better stabilizes topic assignment [27]. LDA can also be built on a conditional random field, two-layer bidirectional long short-term memory, or other neural network representations [23, 28, 29]. Although LDA is traditionally used for longer documents such as news articles and academic papers, LDA has also been applied to Twitter where each tweet is considered a document. To address the limitations caused by short texts such as tweets, researchers have used aggregation schemes where tweets by the same author or with the same terms, hashtags, posted time are combined as one document [9, 10, 30, 23]. Zhao et al. have also used LDA to study the differences between Twitter and New York Times in terms of the discussed topics and content [31], while Aiello et al. [32] applied LDA for the purpose of trending topics detection in sports and politics, using different textual pre-processing steps. Similarly, researchers have modified LDA to capture the temporal nature of documents, such as the Topic over Time (TOT) algorithm [33] for detecting topical trends over continuous time, and Temporal-LDA [34] for modelling topics and their transitions in streaming documents. LDA has also been applied in various domains, such as urban analytics [35, 36], advertising/marketting [37, 38], diseases/medical [39, 40], climate sentiment measurement [41], communication research [42], and aspect-based product review [43].

2.3 Social Media Analytics

Apart from studying topics on microblogs, topic models have also been used to enhance other tasks such as distinguishing between personal and corporate accounts [44] and identifying fake follower accounts [45]. Contrasting opinion topic models find opinions from multiple perspectives in news media [46]. Stances on different opinions can further be used to detect disinformation [47, 48], analyze sentiment to improve the stock prediction [49], or study correlation between topics and their prevalence [50, 51, 52] as a social science task. Topic models are helpful in recommendation systems. Topic models for software similarity [53, 54] help in recommending suitable open-source software repositories for developers [55]. Modelling travellers' preference *e.g.*, cultural, city, or landmark, from the textual description of photos [56, 57] can help travelling recommendation. Similar travellers could be identified according to similar topic preferences. Moreover, modelling textual descriptions of photos could in turn help recognize images on social media [58] as well. In addition to direct application, topic models can assist other algorithms to solve more tasks, such as news or legal document summarization [59, 60].

2.4 Discussion

These earlier related works conducted various studies and provided interesting insights into their applications and main findings of topic models on microblogs such as Twitter. In addition, they have also proposed various novel topic modelling algorithms that have shown good performance on different types of datasets, particularly short texts such as microblogs. Building upon these works, our research and proposed method differ from these earlier works in the following ways:

1 In contrast to previous works that study discussion topics on microblogs, these works approach this problem by applying topic modelling algorithms on microblogs with the aim of understanding topical trends in the microblogging community from a social perspective. These earlier works focus less on classifying individual tweets into specific topics and as a result, they do not emphasise on the performance evaluation on these algorithms.

2 While there are researchers that employ community detection algorithms for understanding discussion topics, we perform an empirical study based on an extensive range of network types (with multiple definitions of vertices and edges), instead of using only word co-occurrence. In addition to a standard word co-occurrence network, we also experiment and evaluate a variant of ClusTop that utilizes word embeddings and the distance similarity between the works for community detection and deriving the discussion topics. In terms of experimental evaluation, we also focus on validating the performance of our proposed algorithm on a set of labelled tweets, instead of only understanding the broad topical trends.

3 Although existing topic models have been adapted to microblogs and short texts with relatively good performance, these algorithms typically require the tuning and setting of appropriate values for various algorithmic parameters, such as the number of topics to model and the Dirichlet prior for both document–topic distributions and topic–word distributions, which are modelled by the k, *alpha* and *beta* parameters, respectively. In contrast, our ClusTop algorithm automatically determines the number of topics and does not require any parameter to be set, due to its local maximization of modularity.

3 Proposed Algorithm

We now describe our proposed algorithm by first defining the basic notations and preliminaries used in this algorithm. Using standard network theory notations, we denote V and E to represent the set of vertices and edges, respectively. Following this, an undirected graph $G = (V, E)$ is represented as a collection of vertices V that are connected by a set of edges E. In turn, each edge $e \in E$ is denoted by $e = (\{v_i, v_j\}, w)$, where w represents the weight of the link between vertices v_i and v_j. In our application of community detection algorithms to topic modelling, we first explore the use of an undirected graph as $G = (U, R)$, where U is the set of unigrams (vertices) and R is the set of relations (edges) between the unigrams. In later sections of this paper, we further examine the effects of different definitions of vertices, such as bi-grams, tri-grams, hashtags, etc, as well as different types of edge weights, such as frequency counts and similarity distances based on word embedding.

In this work, we propose the Clustering-based Topic Modelling (ClusTop) algorithm that uses community detection approaches to topic modelling, based on the undirected graph $G = (U, R)$ and different definitions of unigrams and relations. We first provide an overview of the basic ClusTop algorithm, which consists of the following steps:

1 **Network Construction**. The first step of this algorithm involves constructing a unigram network, *i.e.*, an undirected graph $G = (U, R)$, based on a particular definition of vertices (unigrams) and edges (relations). This step will be elaborated further in Section 3.1, where we will describe the various types of vertices (unigrams) and edges (relations) modelled in this work.

2 **Community Detection**. Using the network graph obtained from Step 1, we will next apply community detection approaches to identify the main communities (topics), where sets of vertices (unigrams) will be grouped into different communities that represent different topics. This step will be further described in Section 3.2.

3 **Topic Assignment**. Based on the detected community from Step 2 that corresponds to a specific topic, this step examines individual tweets and aims to assign this tweet to a specific community. In short, this step aims to label each tweet with an appropriate topic. More details about this step are provided later in Section 3.3.

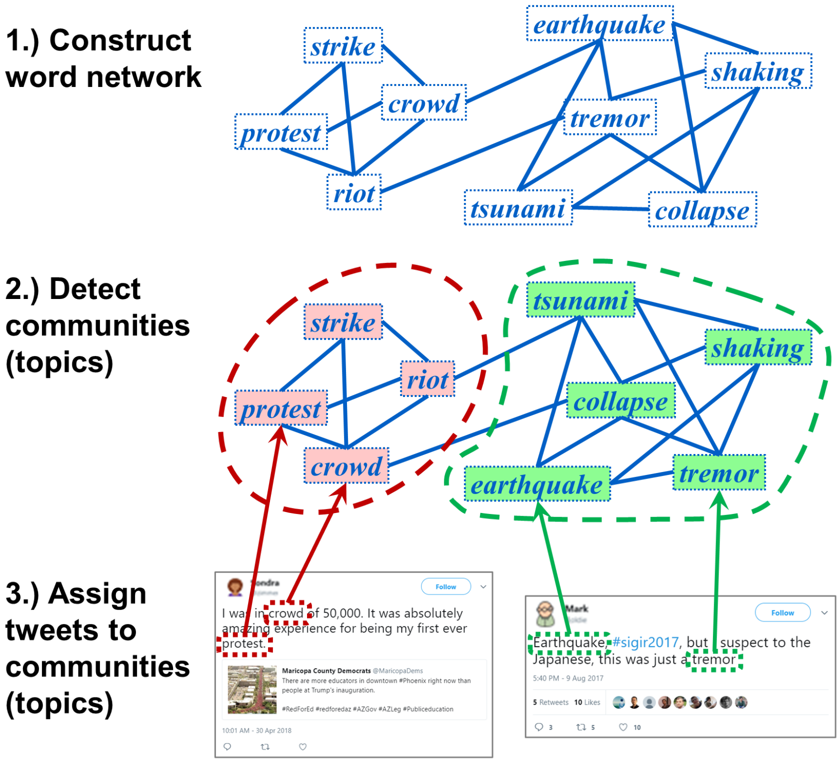
Figure 1 provides an overview of our ClusTop algorithm, with the three main steps of network construction, community (topic) detection and topic assignment. In our subsequent experiments, the steps of network construction and community detection are performed only on the training set, while topic assignment is performed and evaluated on the testing set.

3.1 Network Construction

In this section, we will describe the first step of network graph construction. There are different ways of constructing this network graph, which depends on: (i) the type of network based on different definitions of unigrams (vertices) and their relations (edges); and (ii) the type of document aggregation, *i.e.*, individual tweets, aggregated by hashtag or mentions.

3.1.1 Types of Network

The first stage of our algorithm involves constructing a network graph of word usage, as shown in Algorithm 1. This algorithm involves the following: (i) examining all tweets and tokenizing all words in each tweet based on whitespaces; (ii) for each word-pair in each tweet, build a weighted edge e linking the two words; and (iii) repeating Steps 1 and 2 for all tweets, until we obtain a network graph, where the vertices represent uni-grams and edges represent a relation between two unigrams.



I was in crowd of 50,000. It was absolutely amazing experience for being my first ever protest.

Maricopa County Democrats @MaricopaDems
There are more educators in downtown #Phoenix right now than people at Trump's inauguration.

#RedForEd #redforedaz #AZGov #AZLeg #Publiceducation

10:01 AM - 30 Apr 2018

Earthquake, #sigir2017, but I suspect to the Japanese, this was just a tremor

5:40 PM - 9 Aug 2017

5 Retweets 10 Likes

3 5 10

Figure 1 Overview of the ClusTop Algorithm

The choice of vertices and edges will lead to a different type of network graph being constructed. To better examine the effect of these vertices and edges on the graph type, we experiment with a variety of relations types between different types of uni-gram, including the following:

- **Co-word Usage (Word)**. A relationship where two words (uni-grams) are used in the same tweet. That is, co-word usage models all pair-wise word co-occurrence in a tweet, regardless of where the word appeared.
- **Co-hashtag Usage (Hash)**. A relationship where two hashtags are used in the same tweet. Twitter users typically use hashtags to categorize their tweets into themes and topics [61, 62], and thus serve as a suitable form of unigram relation.
- **Co-noun Usage (Noun)**. A relationship where two nouns are used in the same tweet. For determining the noun in a tweet, we utilize the part-of-speech tagging component from Apache OpenNLP library [63], which has been used by many researchers for similar natural language processing [64, 65, 66].
- **Bigram occurrence (BiG)**. A relationship for two words of each bigram in the tweet. Unlike the co-word usage, this bigram occurrence only considers a relation/edge between two words if they are used one after another in sequence.
- **Trigram occurrence (TriG)**. Similar to the earlier bigram occurrence, except that we model a relationship between three words in a trigram, *i.e.*, there is an additional edge between the first and third word.

Algorithm 1: Network Construction

```
input  : $T$: Collection of tweets in corpus.
output: $G = (U, R)$: Network graph of unigrams (vertices) and relations (edges).
begin
    Initialize an empty graph $G$;
    for each tweet $t \in T$ do
        for each word-pair $(p_1, p_2) \in t$ do
            $e \leftarrow (\{p_1, p_2\}, 1)$;
            if edge $e$ exists in graph $G$ then
                increment edge $e$ in graph $G$ by 1;
            else
                add edge $e$ to graph $G$;
    Return $G$;
```

- **Bigram + Hashtag (BiHa)**. A combination of bigram occurrence and co-hashtag usage, we consider each bigram occurrence and add a relation/edge between each word of the bigram and all hashtags in the tweet.

In the above examples, we determine edge weights based on the co-occurrence frequency of terms observed in a set of tweets, *i.e.*, our training set. We also make use of word embedding to model edge weights as the cosine similarity between a pair of words, *i.e.*, more similar words will be linked with a higher edge weight. For this purpose, we first use the GloVe algorithm [67] for generating the word vector based on hashtags used, then construct a network with vertices based on hashtags and edge weights based on the cosine similarity scores between hashtags. In addition to GloVe, we also generalize this variant using other popular word embedding algorithms such as Word2Vec[68] and FastText[69] to better examine the effects of different word embedding techniques on our approach.

We denote the three variants of these word embedding based networks with its similarity based edge weights as:

- **Hash2Vec-Glove (H2VG)**. A network based on co-usage of hashtags in the same tweet, where the edge weights are based on cosine similarity scores of a word vector trained using GloVe [67].
- **Hash2Vec-Word2Vec (H2VW)**. A network based on co-usage of hashtags in the same tweet, where the edge weights are based on cosine similarity scores of a word vector trained using Word2Vec [68].
- **Hash2Vec-FastText (H2VF)**. A network based on co-usage of hashtags in the same tweet, where the edge weights are based on cosine similarity scores of a word vector trained using FastText [69].

3.1.2 Types of Document Aggregation

In the above examples, we are modelling each tweet as a single document for topic modelling purposes. In more traditional topic modelling, each document typically corresponds to a lengthy piece of text (such as a news article, website or abstract) and traditional topic modelling algorithms work better for these types of lengthy documents. Comparatively on Twitter, each document typically corresponds to a much shorter document in the form of a tweet with up to 280 characters. Researchers have found that aggregating multiple tweets into a single document improves the performance of LDA on Twitter [9, 10]. Building upon these findings, we also experiment with different forms of document aggregation scheme for our ClusTop algorithm, including:

- **No Aggregation, *i.e.*, individual tweets (NA)**. The basic representation where each tweet is considered a single document, *i.e.*, no aggregation as per traditional topic modelling.
- **Aggregate by Hashtags (AH)**. Each document comprises a set of tweets that are aggregated based on common hashtags used.
- **Aggregate by Mentions (AM)**. Each document comprises a set of tweets that are aggregated based on common mentions of Twitter users.

3.2 Community Detection

After constructing the network graph in the previous section, we now describe our approach to modelling the topics in this graph using community detection approaches. Our main example in this paper is on the adaptation of the Louvain algorithm [70] for this purpose, as the Louvain algorithm has been shown to be one of the best performing algorithm in a comprehensive survey of community detection algorithms [71].[2]

Our adaptation of the Louvain algorithm [70] for the purpose of topic modelling is described by the pseudo-code in Algorithm 2, which comprises the following steps:

1. Initially, each unigram is placed in its own community/cluster (Line 2).
2. Following which, for each unigram, we examine each neighbour of this unigram and combine two unigrams into the same community/cluster if their modularity gain is the greatest among all of the neighbours (Lines 4 to 16).
3. Next, we build a new network graph where unigrams in the same community/cluster are combined as a single vertex (unigram), and Step 2 is repeated until the modularity score is maximized (Lines 17 to 20).

One of the reasons for the Louvain algorithm's good performance is due to its local adjustment of unigrams (vertices) into communities/clusters, by maximizing the gain in the following modularity function [70]:

$$
\begin{aligned}
Q = {} & \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \\
& \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right]
\end{aligned}
\tag{1}
$$

[2] Using the earlier generated network graph, our approach can also be easily generalized and used as an input to other community detection algorithms. We have also experimented with other popular community detection algorithms such as the Infomap [72] and Label Propagation [73] algorithms. However, the results show that these algorithms have a tendency to generate a large number (hundreds to thousands) of small communities, thus making it unfeasible for our topic modelling purpose. As such, we utilize the Louvain algorithm due to its good performance, compared to these other community detection algorithms.

Algorithm 2: Topic Modelling using Louvain

```
   input  : $G = (U, R)$: Network graph of unigrams (vertices) and relations (edges).
   output: $A = (U, C)$: Assignment of unigrams (vertices) $U$ into communities $C$.
 1 begin
 2     Assign all unigrams $u$ into their own community;
 3     repeat
 4         for each unigram $u \in U$ do
 5             $MaxModularity \leftarrow -1$;
 6             $MaxModNeighbour \leftarrow NULL$;
 7             for each neighbour $u_n$ of unigram $u$ do
 8                 $ShiftMod \leftarrow$ Modularity score of shifting unigram $u$ to neighbour $u_n$'s
                     community;
 9                 if $ShiftMod > MaxModularity$ then
10                     $MaxModularity \leftarrow ShiftMod$;
11                     $MaxModNeighbour \leftarrow u_n$;
12             $OriginalMod \leftarrow$ Modularity score of unigram $u$ in its original community;
13             if $OriginalMod > MaxModularity$ then
14                 Shift unigram $u$ to the community of MaxModNeighbour;
15             else
16                 Keep unigram $u$ in its original community;
17         $U_n \leftarrow$ New unigrams (vertices) $U_n$ based on the newly-formed communities;
18         $R_n \leftarrow$ New relations (edges) $R_n$ based on edge weights between nodes in two
             communities;
19         $G_n \leftarrow$ New network graph $G_n = (U_n, R_n)$;
20         The algorithm iterates again (Lines 4 to 19) with network graph $G_n$ as input;
21     until Community structure stabilizes and modularity score is maximized;
22     Return $a_n$;
```

where $\sum_{in}$ and $\sum_{tot}$ represents the total weight of all links inside a community/cluster and total weight of all links to a community/cluster, respectively. Similarly, the terms k_i and $k_{i,in}$ denote the total weight of all links to i and total weight of links to i within the community/cluster. Lastly, m denotes the total weight of all links in the network graph.

At the end of this step, we will obtain a set of communities/clusters based on the provided network graph. Each community/cluster will represent a particular topic, where the members of each community/cluster serve as the representative words of each topic. For each topic, we also rank the keywords (*i.e.*, members of each community) based on the total weight of all links to a unigram/vertex.

3.3 Topic Assignment

Given the detected communities/topics C from Section 3.2 and a tweet $t = \{u_1, ...u_n\}$, we define the most likely topic for this tweet as:

$$\arg\max_{c \in C} \sum_{u \in c} k_u \delta(u = u_t), \qquad \forall\ u_t \in t \tag{2}$$

where $\delta(u = u_t) = 1$ if a unigram u of a community/topic $c \in C$ is the same as a unigram u_t of a tweet t and $\delta(p = c) = 0$ otherwise, and k_u denotes the total weight of links to unigram u (as previously described in Section 3.2).

In short, we assign a tweet t to a community/topic c that has the highest co-occurrence of unigrams in both the tweet and community/topic, where the unigram in the community/topic is weighted based on its co-occurrence to other unigrams.

4 Dataset and Evaluation Methodology

In this section, we give an overview of our experimental dataset and describe our evaluation methodology in terms of the ClusTop algorithm variants, baseline algorithms and evaluation metrics.

4.1 Dataset

For our experimental evaluation, we utilize three Twitter datasets with labelled topics [74, 75, 76], which enables us to better evaluate our algorithm and baselines against the ground truth topics compared to an unlabelled dataset. In total, these datasets comprise close to 8 million tweets, from which we focus on the subset of tweets with annotated and verified topics. These topics are in the form of 60k labelled tweets about 6 crises [74], 27.9k labelled tweets about 26 crises [75], and 3.6k labelled tweets about 8 events [76]. Refer to Table 1 for more details. The annotation of these tweets into the respective topics (crises and events) were performed via the CrowdFlower crowdsourcing platform, and more details can be found in the respective papers.

Table 1 Description of Dataset

Dataset	Paper Reference	Number of Topics	Total Tweets	Labelled Tweets
A	[74]	6	7.67mil	60k
B	[75]	26	0.28mil	27.9k
C	[76]	8	4.8k	3.6k

We split each dataset into four partitions and perform a 4-fold cross validation [77]. At each evaluation iteration, we use three partitions as our training set and the last partition as our testing set. After completing all evaluations, we compute and report the mean results for each algorithm based on the metrics of topic coherence, pointwise mutual information, precision, recall and f-score, which we elaborate further in the rest of the paper.

4.1.1 Topic Quality Metrics

For determining the quality of the detected topics, we measure the topic quality based on the topic coherence and pointwise mutual information metrics. These two metrics have also been widely used by many topic modelling researchers [78, 79, 10]. For both evaluation metrics, we denote a detected topic t that comprises a set of n representative unigrams/keywords $U^{(t)} = (u_1^{(t)}, ..., u_n^{(t)})$ for each topic.

1 **Topic Coherence (TC)**. Given that $D(u_i, u_j)$ denotes the number of times both unigrams u_i and u_j appeared in the same document/tweet, and similarly, $D(u_i)$ for a single unigram u_i, topic coherence is defined as:

$$TC(t, U^{(t)}) = \sum_{u_i \in U^{(t)}} \sum_{u_j \in U^{(t)}, u_i \neq u_j} log \, \frac{D(u_i, u_j)}{D(u_j)} \qquad (3)$$

2 **Pointwise Mutual Information (PMI)**. Given that $P(u_i, u_j)$ denotes the probability of a unigram pair u_i and u_j appearing in the same document/tweet, and $P(u_i)$ for the probability of a single unigram u_i, pointwise mutual information is defined as:

$$PMI(t, U^{(t)}) = \sum_{u_i \in U^{(t)}} \sum_{u_j \in U^{(t)}, u_i \neq u_j} log \, \frac{P(u_i, u_j)}{P(u_i)P(u_j)} \tag{4}$$

In both the TC and PMI metrics, it is possible for a division by 0 or taking the log of 0 when the appropriate numerator or denominator is 0, *i.e.*, when a particular word or word pair has not been previously observed. As such, we adopt a similar strategy as [78, 10] by adding a small value $\epsilon = 1$ to both components to avoid the situation of a division by 0 or log of 0.

4.1.2 Topic Relevance Metrics

Precision, recall and f-score are popular metrics used in Information Retrieval and other related fields, such as in topic modelling [32, 80], tour recommendation [81, 82, 83], location prediction and tagging [84, 85, 86], event detection [87, 88], among others. In contrast to the previous topic quality metrics (TC and PMI), these metrics allow us to evaluate how relevant and accurate the detected topics are, compared to the ground truth topics. In topic modelling, researchers typically manually curate a set of ground truth keywords to describe a specific topic, then evaluate how well the detected keywords from their topic models match these ground truth keywords [32]. For our evaluation, we adopt a similar methodology except that we automatically determine the ground truth keywords from the respective Wikipedia article for each topic.

Given that $U^D = (u_1^D, ..., u_n^D)$ and $U^G = (u_1^G, ..., u_n^G)$ denotes the set of detected unigrams and ground truth unigrams for a specific topic, the metrics we use are as follows:

- **Precision**. The proportion of unigrams for the detected topic U^D that also appears in the ground truth unigrams U^G. For a topic t, precision is defined as:

$$P(t) = \frac{|U^D \cap U^G|}{|U^D|} \tag{5}$$

- **Recall**. The proportion of ground truth unigrams U^G that also appears in the unigrams for the detected topic U^D. For a topic t, recall is defined as:

$$R(t) = \frac{|U^D \cap U^G|}{|U^G|} \tag{6}$$

- **F-score**. The harmonic mean of precision $P(t)$ and recall $R(t)$, which was introduced in Equations 5 and 6, respectively. For a topic t, F-score is defined as:

$$F(t) = \frac{2 \times P(t) \times R(t)}{P(t) + R(t)} \tag{7}$$

In our experiments, we compute the precision, recall and F-score derived from the testing set, in terms of the top 5 and 10 keywords of each topic modelled.

4.1.3 Summary Rank Metrics

As our experiments involve five evaluation metrics, three datasets and 18 algorithms, we develop an intuitive approach to represent the performance of each algorithm. This approach first ranks an algorithm's performance from 1 to 18 for each evaluation metric and dataset, with the lowest rank being the best performing one. For each combination of topic quality metrics (topic coherence and pointwise mutual information) and topic relevance metrics (precision, recall and f-score), we take the average of each metrics group across all three datasets for an average rank. For example, if an algorithm ranked 1st, 1st and 2nd in terms of topic coherence and 2nd, 1st, 2nd in terms of pointwise mutual information for datasets A, B, C, respectively, this algorithm will be assigned an overall rank of 1.5 for the topic quality metric.

4.2 Variants of ClusTop Algorithm

Based on the six types of unigram network and three types of document aggregation (introduced in Section 3.1), there can be mutliple variants of our ClusTop algorithm. For our evaluation, we experiment with the following 21 variants of our ClusTop algorithm, namely:

- **ClusTop-Word-NA**. ClusTop based on a co-word usage network, with no tweet aggregation.
- **ClusTop-BiG-NA**. ClusTop based on a bigram occurrence network, with no tweet aggregation.
- **ClusTop-TriG-NA**. ClusTop based on a trigram occurrence network, with no tweet aggregation.
- **ClusTop-BiHa-NA**. ClusTop based on a bigram occurrence + co-hashtag usage network, with no tweet aggregation.
- **ClusTop-Hash-NA**. ClusTop based on a co-hashtag usage network, with no tweet aggregation.
- **ClusTop-H2VG-NA**. ClusTop based on a co-hashtag usage network where edge weights are based on hash2vec-GloVe scores between hashtags, with no tweet aggregation.
- **ClusTop-H2VW-NA**. ClusTop based on a co-hashtag usage network where edge weights are based on hash2vec-Word2Vec scores between hashtags, with no tweet aggregation.
- **ClusTop-H2VF-NA**. ClusTop based on a co-hashtag usage network where edge weights are based on hash2vec-FastText scores between hashtags, with no tweet aggregation.
- **ClusTop-Noun-NA**. ClusTop based on a co-noun usage network, with no tweet aggregation.

- **ClusTop-Word-AH**. ClusTop based on a co-word usage network, with tweets aggregated based on common hashtags.
- **ClusTop-Hash-AH**. ClusTop based on a co-hashtag usage network, with tweets aggregated based on common hashtags.
- **ClusTop-H2VG-AH**. ClusTop based on a co-hashtag usage network where edge weights are based on hash2vec-GloVe scores between hashtags, with tweets aggregated based on common hashtags.
- **ClusTop-H2VW-AH**. ClusTop based on a co-hashtag usage network where edge weights are based on hash2vec-Word2Vec scores between hashtags, with tweets aggregated based on common hashtags.
- **ClusTop-H2VF-AH**. ClusTop based on a co-hashtag usage network where edge weights are based on hash2vec-FastText scores between hashtags, with tweets aggregated based on common hashtags.
- **ClusTop-Noun-AH**. ClusTop based on a co-noun usage network, with tweets aggregated based on common hashtags.
- **ClusTop-Word-AM**. ClusTop based on a co-word usage network, with tweets aggregated based on common mentions.
- **ClusTop-Hash-AM**. ClusTop based on a co-hashtag usage network, with tweets aggregated based on common mentions.
- **ClusTop-H2VG-AM**. ClusTop based on a co-hashtag usage network where edge weights are based on hash2vec-GloVe scores between hashtags, with tweets aggregated based on common mentions.
- **ClusTop-H2VW-AM**. ClusTop based on a co-hashtag usage network where edge weights are based on hash2vec-Word2Vec scores between hashtags, with tweets aggregated based on common mentions.
- **ClusTop-H2VF-AM**. ClusTop based on a co-hashtag usage network where edge weights are based on hash2vec-FastText scores between hashtags, with tweets aggregated based on common mentions.
- **ClusTop-Noun-AM**. ClusTop based on a co-noun usage network, with tweets aggregated based on common mentions.

Note that we did not use the ClusTop variants based on bigrams and trigrams combined with the hashtag and mention aggregation schemes, as these variants provide minimal improvements compared to their original non-aggregated variants. Consider a simple example of three tweets with a common hashtag, the hashtag aggregation scheme with bigrams will only produce an additional two bigrams resulting from the first and second tweet as well as the second and third tweet. Moreover, these two additional bigrams will be generated from the last word of the first tweet and the first word of the second tweet, which will not be syntactically meaningful in most cases.

4.3 Baseline Algorithms

LDA is a popular topic modelling algorithm that was used for traditional documents (such as news articles), and more recently for social media (such as tweets on Twitter). Given the popularity of LDA for topic modelling, we compare our ClusTop algorithm and its variants against the following LDA-based algorithms, namely:

1 **LDA-Orig**. The original version of LDA introduced by [6], where each document corresponds to a single tweet.

Table 2 Comparison of ClusTop algorithm against various baselines, in terms of Topic Coherence (TC) and Pointwise Mutual Information (PMI) for the top 5, 10, 15 and 20 keywords. The rank of an algorithm's performance for each metric are provided in brackets.

Algorithm	Rank@5	Rank@10	Rank@15	Rank@20
ClusTop-Word-NA	(16.7)	(17.7)	(18.7)	(19.2)
ClusTop-BiG-NA	(15.7)	(16.8)	(17.0)	(17.3)
ClusTop-TriG-NA	(15.2)	(16.7)	(17.2)	(17.0)
ClusTop-BiHa-NA	(11.0)	(12.3)	(14.3)	(14.5)
ClusTop-Hash-NA	(6.0)	(5.3)	(4.8)	(5.7)
ClusTop-Noun-NA	(7.5)	(9.2)	(10.5)	(10.3)
ClusTop-H2VG-NA	(11.5)	(10.7)	(9.2)	(9.0)
ClusTop-H2VW-NA	(2.0)	(1.8)	(1.8)	(2.0)
ClusTop-H2VF-NA	(2.0)	(2.0)	(1.8)	(2.0)
ClusTop-Word-AH	(13.8)	(15.3)	(16.8)	(17.8)
ClusTop-Hash-AH	(5.5)	(5.2)	(4.7)	(4.5)
ClusTop-Noun-AH	(13.2)	(15.2)	(15.7)	(16.0)
ClusTop-H2VG-AH	(13.8)	(14.2)	(13.5)	(12.5)
ClusTop-H2VW-AH	(17.5)	(15.3)	(14.0)	(12.8)
ClusTop-H2VF-AH	(18.3)	(16.8)	(16.2)	(16.0)
ClusTop-Word-AM	(12.5)	(14.2)	(15.0)	(15.8)
ClusTop-Hash-AM	(8.3)	(8.2)	(9.3)	(9.8)
ClusTop-Noun-AM	(8.0)	(7.3)	(6.7)	(6.5)
ClusTop-H2VG-AM	(15.0)	(14.3)	(13.2)	(11.8)
ClusTop-H2VW-AM	(9.2)	(6.5)	(5.0)	(5.0)
ClusTop-H2VF-AM	(8.0)	(6.0)	(5.7)	(5.3)
LDA-Orig	(24.0)	(24.0)	(24.0)	(24.0)
LDA-Hash	(21.8)	(22.0)	(22.0)	(22.0)
LDA-Ment	(22.8)	(23.0)	(23.0)	(23.0)

2 **LDA-Hash**. A variant of LDA applied on Twitter, where each document is aggregated from multiple tweets with the same hashtag [10].

3 **LDA-Ment**. An adaptation of the Twitter-based LDA variant proposed by [89], where we aggregate tweets with the same mention into a single document.

5 Experimental Results and Discussion

In this section, we report on the results of our experiments and discuss some implications of these findings.

5.1 Topic Coherence and Pointwise Mutual Information

Table 2 shows a summary of the performance of our ClusTop algorithm and its variants against the various LDA baselines, in terms of average rank based on Topic Coherence and Pointwise Mutual Information scores on the top 5, 10, 15 and 20 keywords in the detected topics. For a more detailed breakdown, Tables 4 and 5 show the performance of our ClusTop algorithm and its variants against the various LDA baselines, in terms of Topic Coherence and Pointwise Mutual Information, based on the top 5, 10, 15 and 20 keywords in the detected topics.

The results generally show that all variants of our ClusTop algorithm outperform the various LDA baselines, in terms of the average rank metrics. All ClusTop variants also out-perform the LDA baselines in terms of the individual evaluation metrics of Topic Coherence and Pointwise Mutual Information across all datasets. In particular, we note the following:

- The performance of ClusTop could be largely attributed to its usage of the various types of word network graphs, which retain the syntactic meaning and association between words in a tweet via the use of vertices in the form of unigrams, bigrams, trigrams and its variants, and edges in the form of word co-occurrence usage and various types of word embedding similarity distances.
- All ClusTop variants that utilize hashtags (ClusTop-Hash-NA, ClusTop-Hash-AH and ClusTop-Hash-AM) offer better overall performance compared to its counterparts that utilizes other forms of unigram and relation, *i.e.*, words, bigrams, trigrams, nouns.
- The aggregation schemes employed by LDA (LDA-Hash and LDA-Ment) generally outperform their original counterpart (LDA-Orig), thus showing that LDA works better on larger documents.
- In addition to all ClusTop variants outperforming the LDA baselines, the aggregation schemes employed by ClusTop showed better performance compared to their non-aggregated counterparts.

5.2 Precision, Recall and F-score

Table 3 shows the average ranks based on the Precision, Recall and F-score scores of our ClusTop algorithm and variants, and the various LDA baselines based on the top 5, 10, 15 and 20 keywords of detected topics. For a more detailed breakdown of the results, Table 6 shows the Precision, Recall and F-score of our ClusTop algorithm and variants, and the various LDA baselines based on the top 5 keywords of detected topics, while Tables 7, 8, and 9 show the same results based on top 10, 15 and 20 keywords of detected topics, respectively.

There are specific variants of ClusTop that outperform the LDA baselines in terms of Precision, Recall and F-score. Our main observations are as follows:

- In terms of overall rank (average of Precision, Recall and F-score), ClusTop-BiG-NA, ClusTop-TriG-NA, ClusTop-H2VG-AM offers the best overall performance.
- In terms of precision, the **Hash2Vec** variants consistently fosters the best performers (except one case in Table 6 where LDA-Orig is ranked 2nd, beating all **Hash2vec** variants except ClusTop-H2VW-AH). The next two best performers are ClusTop-Hash-NA and ClusTop-Noun-AM, except for the aforementioned case where they are beaten by LDA-Orig Table 6.
- ClusTop-H2VW-* and ClusTop-H2VF-* have slightly higher precision than ClusTop-H2VG-*, observed from that ClusTop-H2VG-* never scores top precision in Table 6. The existence of this slight difference is due to each word embedding algorithm being trained using its own vocabulary set. While GloVe (twitter.27B) provides embeddings for some hashtags that are not standard English, the other two algorithms do not. Missing these hashtags possibly increases precision of ClusTop-H2VW-* and ClusTop-H2VF-*

Table 3 Comparison of ClusTop algorithm against various baselines, in terms of Precision (Pre), Recall (Rec) and F-score (FS) for the top 5 keywords/unigrams of each topic. The rank of an algorithm's performance for each metric are provided in brackets.

Algorithm	Rank@5	Rank@10	Rank@15	Rank@20
ClusTop-Word-NA	(10.8)	(10.9)	(10.6)	(10.4)
ClusTop-BiG-NA	(8.3)	(8.6)	(8.1)	(8.3)
ClusTop-TriG-NA	(7.7)	(8.3)	(8.2)	(8.7)
ClusTop-BiHa-NA	(9.3)	(9.3)	(9.8)	(9.4)
ClusTop-Hash-NA	(15.4)	(15.1)	(15.1)	(15.0)
ClusTop-Noun-NA	(9.8)	(9.2)	(10.1)	(10.2)
ClusTop-H2VG-NA	(12.1)	(12.2)	(12.7)	(12.6)
ClusTop-H2VW-NA	(16.1)	(15.0)	(15.6)	(15.3)
ClusTop-H2VF-NA	(15.3)	(15.6)	(14.9)	(14.9)
ClusTop-Word-AH	(13.8)	(14.3)	(15.6)	(13.9)
ClusTop-Hash-AH	(10.2)	(12.3)	(12.1)	(11.7)
ClusTop-Noun-AH	(13.0)	(13.8)	(13.2)	(13.0)
ClusTop-H2VG-AH	(10.8)	(10.3)	(11.8)	(12.0)
ClusTop-H2VW-AH	(14.8)	(16.2)	(15.3)	(15.8)
ClusTop-H2VF-AH	(15.1)	(14.2)	(14.0)	(14.3)
ClusTop-Word-AM	(13.8)	(13.8)	(11.2)	(10.8)
ClusTop-Hash-AM	(13.0)	(11.1)	(10.6)	(10.7)
ClusTop-Noun-AM	(12.7)	(12.9)	(13.6)	(13.7)
ClusTop-H2VG-AM	(9.1)	(10.7)	(11.1)	(11.1)
ClusTop-H2VW-AM	(15.4)	(14.0)	(15.1)	(15.1)
ClusTop-H2VF-AM	(16.6)	(15.9)	(16.0)	(16.2)
LDA-Orig	(9.7)	(12.0)	(11.4)	(11.6)
LDA-Hash	(11.2)	(9.4)	(8.8)	(9.0)
LDA-Ment	(12.4)	(11.4)	(11.0)	(11.9)

in Table 6, as vast majority of ground truth words are in standard English vocabulary.

6 Conclusion and Future Work

In this paper, we proposed the ClusTop algorithm for topic modelling on Twitter, using community detection approaches on a network graph with multiple definitions of vertices and edges. While traditional topic modelling algorithms require the tuning and setting of numerous parameters, ClusTop does not require this parameter tuning and is able to automatically determine the appropriate number of topics using a local maximization of modularity among the word network graph. We also performed an empirical study on the effects of using different types of vertices (unigrams, bigrams, trigrams, hashtags, nouns from part-of-speech tagging) types of edges (word co-occurrence frequency and word embedding similarity distance), and different aggregation schemes (individual tweets, hashtags and mentions). The different possible combinations of vertices, edges and aggregation schemes results in multiple variants of our ClusTop algorithm, which we use to compare among the variants as well as against various LDA baselines. Our experimental evaluation on the ClusTop variants and baselines are based on the evaluation metrics of topic coherence, pointwise mutual information, precision, recall and F-score. The experimental results based on three Twitter datasets with labeled topics (crises and events) show that our ClusTop algorithm out-performs the various LDA baselines in terms of these evaluation metrics.

This work explored how community detection approaches alongside different types of word network graphs can be used for automated topic modelling on Twitter. We performed an empirical study to examine the effects of different types of network graphs based on different definitions of vertices, edges and aggregation schemes on a variety of performance metrics. There still remain various directions for future research, which include:

- A major challenge in evaluating topic models and text classification models is the requirement of a dataset with annotated labels of the ground truth topics. A possible future direction is to automate the labelling of this ground truth topic by using the semantic similarity between tweets or other short texts and Wikipedia or news articles to assign the appropriate topic labels based on the categorisation for the latter.
- Our work is primarily focused on using community detection approaches for topic modelling purposes and does not incorporate other aspects of a social network, such as friendship links. Future work can utilize a joint modelling of social relations between users and the various types of word network graph to detect topic-coherence communities, *i.e.*, communities of users based on topical interests.
- Another future direction is to extend our ClusTop algorithm to incorporate temporal and spatial attributes associated with geo-tagged tweets. With the increased use of smart devices and geo-tagged social media, this consideration of temporal and spatial attributes will enable researchers to better model topics that are associated with specific time periods or physical locations.

Ethics approval and consent to participate
Not applicable.

Consent for publication
The authors consent to the publication of this manuscript.

Availability of data and materials
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Funding
This research is funded by MyIP:7293 and SRG-ISTD-2018-140.

Author's contributions
WM and KHL designed the research, ran experiments, analyzed results, and wrote the manuscript. JL, SK, LF and AH designed the research, analyzed results and contributed to manuscript preparation. All authors read and approved the final manuscript.

Acknowledgements
This research is funded in part by the Defence Science and Technology Group, Edinburgh, South Australia, under contract MyIP:7293, and the Singapore University of Technology and Design under grant SRG-ISTD-2018-140.

Author details
[1]Engineering Product Development Pillar, Singapore University of Technology and Design, Singapore. [2]Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore. [3]Forth AI, Singapore. [4]School of Computing and Information Systems, The University of Melbourne, Australia.

References

1. Statistics, I.L.: Twitter Usage Statistics. Internet. http://www.internetlivestats.com/twitter-statistics/ (2016)
2. Kumar, S., Morstatter, F., Liu, H.: Twitter Data Analytics. Springer, New York, NY, USA (2013)
3. Liao, Y., Moshtaghi, M., Han, B., Karunasekera, S., Kotagiri, R., Baldwin, T., Harwood, A., Pattison, P.: Mining Micro-Blogs: Opportunities and Challenges. Social Networks: Computational Aspects and Mining. London in the Computer Communications and Networks series. Springer, ??? (2011)
4. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Analysing how people orient to and spread rumours in social media by looking at conversational threads. Journal of the American Society for Information Science **41**(6), 391 (1990)
5. Hofmann, T.: Probabilistic latent semantic analysis. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99), pp. 289–296 (1999)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
7. De Smet, W., Moens, M.-F.: Cross-language linking of news stories on the web using interlingual topic modelling. In: Proceedings of the 2nd ACM Workshop on Social Web Search and Mining, pp. 57–64 (2009)
8. Jacobi, C., Van Atteveldt, W., Welbers, K.: Quantitative analysis of large amounts of journalistic texts using topic modelling. Digital journalism **4**(1), 89–106 (2016)
9. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the First Workshop on Social Media Analytics (SMA'10), pp. 80–88 (2010)
10. 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 (SIGIR'13), pp. 889–892 (2013)
11. Lim, K.H., Karunasekera, S., Harwood, A.: Clustop: A clustering-based topic modelling algorithm for twitter using word networks. In: Proceedings of the 2017 IEEE International Conference on Big Data (BigData'17), pp. 2009–2018 (2017)
12. Liu, Y., Goncalves, J., Ferreira, D., Xiao, B., Hosio, S., Kostakos, V.: CHI 1994-2013: mapping two decades of intellectual progress through co-word analysis. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI'14), pp. 3553–3562 (2014)
13. Lim, K.H., Datta, A.: A topological approach for detecting twitter communities with common interests. In: Ubiquitous Social Media Analysis, pp. 23–43. Springer, ??? (2013)
14. Lim, K.H., Datta, A.: An interaction-based approach to detecting highly interactive twitter communities using tweeting links. Web Intelligence **14**(1), 1–15 (2016)
15. Paranyushkin, D.: Identifying the pathways for meaning circulation using text network analysis. In: Nodus Labs (2011)
16. Jr, S.B., Kido, G.S., Tavares, G.M.: Artificial and natural topic detection in online social networks. iSys - Revista Brasileira de Sistemas de Informacao **10**(1), 80–98 (2017)
17. Fried, D., Surdeanu, M., Kobourov, S., Hingle, M., Bell, D.: Analyzing the language of food on social media. In: Proceedings of the 2014 IEEE International Conference on Big Data (BigData'14), pp. 778–783 (2014)
18. Surian, D., Nguyen, D.Q., Kennedy, G., Johnson, M., Coiera, E., Dunn, A.G.: Characterizing twitter discussions about hpv vaccines using topic modeling and community detection. Journal of Medical Internet Research **18**(8) (2016)
19. Amati, G., Angelini, S., Cruciani, A., Fusco, G., Gaudino, G., Pasquini, D., Vocca, P.: Topic modeling by community detection algorithms. In: Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks, pp. 15–20 (2021)
20. Martin, F., Johnson, M.: More efficient topic modelling through a noun only approach. In: Proceedings of the Australasian Language Technology Association Workshop 2015, Parramatta, Australia, pp. 111–115 (2015). https://aclanthology.org/U15-1013
21. Yang, S., Zhang, H.: Text mining of twitter data using a latent dirichlet allocation topic model and sentiment analysis. Int. J. Comput. Inf. Eng **12**(7), 525–529 (2018)
22. Sun, Y., Loparo, K., Kolacinski, R.: Conversational structure aware and context sensitive topic model for online discussions. In: 2020 IEEE 14th International Conference on Semantic Computing (ICSC), pp. 85–92 (2020). IEEE
23. Gao, W., Peng, M., Wang, H., Zhang, Y., Xie, Q., Tian, G.: Incorporating word embeddings into topic modeling of short text. Knowledge and Information Systems **61**(2), 1123–1145 (2019)
24. Dieng, A.B., Ruiz, F.J., Blei, D.M.: Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics **8**, 439–453 (2020)
25. Dai, X., Bikdash, M., Meyer, B.: From social media to public health surveillance: Word embedding based clustering method for twitter classification. In: SoutheastCon 2017, pp. 1–7 (2017). IEEE
26. Bagheri, A., Sammani, A., van der Heijden, P.G., Asselbergs, F.W., Oberski, D.L.: Etm: Enrichment by topic modeling for automated clinical sentence classification to detect patients' disease history. Journal of Intelligent Information Systems **55**(2), 329–349 (2020)
27. Nikolenko, S.I., Koltcov, S., Koltsova, O.: Topic modelling for qualitative studies. Journal of Information Science **43**(1), 88–102 (2017)
28. Jansson, P., Liu, S.: Distributed representation, LDA topic modelling and deep learning for emerging named entity recognition from social media. In: Proceedings of the 3rd Workshop on Noisy User-generated Text, pp. 154–159. Association for Computational Linguistics, Copenhagen, Denmark (2017). doi:10.18653/v1/W17-4420. https://aclanthology.org/W17-4420
29. Bhat, M.R., Kundroo, M.A., Tarray, T.A., Agarwal, B.: Deep lda: A new way to topic model. Journal of Information and Optimization Sciences **41**(3), 823–834 (2020)
30. Steinskog, A., Therkelsen, J., Gambäck, B.: Twitter topic modeling by tweet aggregation. In: Proceedings of the 21st Nordic Conference on Computational Linguistics, pp. 77–86. Association for Computational

Linguistics, Gothenburg, Sweden (2017). https://aclanthology.org/W17-0210

31. 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: Proceedings of the 33rd European Conference on Information Retrieval (ECIR'11), pp. 338–349 (2011)
32. Aiello, L.M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Göker, A., Kompatsiaris, I., Jaimes, A.: Sensing trending topics in twitter. IEEE Transactions on Multimedia **15**(6), 1268–1282 (2013)
33. Wang, X., McCallum, A.: Topics over time: A non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06), pp. 424–433 (2006)
34. Wang, Y., Agichtein, E., Benzi, M.: Tm-lda: Efficient online modeling of latent topic transitions in social media. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12), pp. 123–131 (2012)
35. Lansley, G., Longley, P.A.: The geography of twitter topics in london. Computers, Environment and Urban Systems **58**, 85–96 (2016)
36. Wang, J., Feng, Y., Naghizade, E., Rashidi, L., Lim, K.H., Lee, K.E.: Happiness is a choice: Sentiment and activity-aware location recommendation. In: Proceedings of the 2018 Web Conference Companion (WWW'18), pp. 1401–1405 (2018)
37. Chen, Y., Amiri, H., Li, Z., Chua, T.-S.: Emerging topic detection for organizations from microblogs. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13), pp. 43–52 (2013)
38. Barry, A.E., Valdez, D., Padon, A.A., Russell, A.M.: Alcohol advertising on twitter—a topic model. American Journal of Health Education **49**(4), 256–263 (2018)
39. Missier, P., Romanovsky, A., Miu, T., Pal, A., Daniilakis, M., Garcia, A., Cedrim, D., da Silva Sousa, L.: Tracking dengue epidemics using twitter content classification and topic modelling. In: Proceedings of the 2016 International Conference on Web Engineering (ICWE'16), pp. 80–92 (2016)
40. Kwan, J.S.-L., Lim, K.H.: Understanding public sentiments, opinions and topics about covid-19 using twitter. In: Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'20), pp. 623–626 (2020)
41. Dahal, B., Kumar, S.A., Li, Z.: Topic modeling and sentiment analysis of global climate change tweets. Social Network Analysis and Mining **9**(1), 1–20 (2019)
42. Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., *et al.*: Applying lda topic modeling in communication research: Toward a valid and reliable methodology. Communication Methods and Measures **12**(2-3), 93–118 (2018)
43. Jeong, B., Yoon, J., Lee, J.-M.: Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. International Journal of Information Management **48**, 280–290 (2019)
44. Yin, P., Ram, N., Lee, W.-C., Tucker, C., Khandelwal, S., Salathe, M.: Two sides of a coin: Separating personal communication and public dissemination accounts in twitter. In: Proceedings of the 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'14), pp. 163–175 (2014)
45. Shen, Y., Yu, J., Dong, K., Nan, K.: Automatic fake followers detection in chinese micro-blogging system. In: Proceedings of the 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'14), pp. 596–607 (2014)
46. Fang, Y., Si, L., Somasundaram, N., Yu, Z.: Mining contrastive opinions on political texts using cross-perspective topic model. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, pp. 63–72 (2012)
47. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter **19**(1), 22–36 (2017)
48. Song, X., Petrak, J., Jiang, Y., Singh, I., Maynard, D., Bontcheva, K.: Classification aware neural topic model for covid-19 disinformation categorisation. PloS one **16**(2), 0247086 (2021)
49. Nguyen, T.H., Shirai, K.: Topic modeling based sentiment analysis on social media for stock market prediction. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1354–1364 (2015)
50. Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D.G.: Structural topic models for open-ended survey responses. American Journal of Political Science **58**(4), 1064–1082 (2014)
51. Roberts, M.E., Stewart, B.M., Airoldi, E.M.: A model of text for experimentation in the social sciences. Journal of the American Statistical Association **111**(515), 988–1003 (2016)
52. Grimmer, J.: A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. Political Analysis **18**(1), 1–35 (2010)
53. Tian, K., Revelle, M., Poshyvanyk, D.: Using latent dirichlet allocation for automatic categorization of software. In: 2009 6th IEEE International Working Conference on Mining Software Repositories, pp. 163–166 (2009). IEEE
54. Linstead, E., Rigor, P., Bajracharya, S., Lopes, C., Baldi, P.: Mining concepts from code with probabilistic topic models. In: Proceedings of the Twenty-second IEEE/ACM International Conference on Automated Software Engineering, pp. 461–464 (2007)
55. Di Rocco, J., Di Ruscio, D., Di Sipio, C., Nguyen, P., Rubei, R.: Topfilter: an approach to recommend relevant github topics. In: Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), pp. 1–11 (2020)
56. Jiang, S., Qian, X., Shen, J., Mei, T.: Travel recommendation via author topic model based collaborative filtering. In: International Conference on Multimedia Modeling, pp. 392–402 (2015). Springer
57. Hu, B., Ester, M.: Spatial topic modeling in online social media for location recommendation. In: Proceedings

of the 7th ACM Conference on Recommender Systems, pp. 25–32 (2013)

58. Niu, Z., Hua, G., Gao, X., Tian, Q.: Semi-supervised relational topic model for weakly annotated image recognition in social media. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4233–4240 (2014)
59. Alguliyev, R.M., Aliguliyev, R.M., Isazade, N.R., Abdi, A., Idris, N.: Cosum: Text summarization based on clustering and optimization. Expert Systems **36**(1), 12340 (2019)
60. Nagwani, N.K.: Summarizing large text collection using topic modeling and clustering based on mapreduce framework. Journal of Big Data **2**(1), 1–18 (2015)
61. Ma, Z., Sun, A., Cong, G.: Will this #hashtag be popular tomorrow? In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12), pp. 1173–1174 (2012)
62. Lehmann, J., Goncalves, B., Ramasco, J.J., Cattuto, C.: Dynamical classes of collective attention in twitter. In: Proceedings of the 21st International Conference on World Wide Web (WWW'12), pp. 251–260 (2012)
63. Foundation, T.A.S.: The Apache OpenNLP library. Internet. http://opennlp.apache.org (2017)
64. Mattmann, C.A., Sharan, M.: An automatic approach for discovering and geocoding locations in domain-specific web data. In: Proceedings of the 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI'16), pp. 87–93 (2016)
65. Vicente, I.S., Saralegi, X., Agerri, R.: Elixa: A modular and flexible absa platform. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15), pp. 748–752 (2015)
66. Agerri, R., Rigau, G.: Robust multilingual named entity recognition with shallow semi-supervised features. Artificial Intelligence **238**, 63–82 (2016)
67. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14), pp. 1532–1543 (2014)
68. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
69. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
70. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics **2008**(10), 10008 (2008)
71. Fortunato, S.: Community detection in graphs. Physics Reports **486**(3), 75–174 (2010)
72. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Science **105**(4), 1118–1123 (2008)
73. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E **76**(3), 036106 (2007)
74. Olteanu, A., Castillo, C., Diaz, F., Vieweg, S.: Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM'14), pp. 376–385 (2014)
75. Olteanu, A., Vieweg, S., Castillo, C.: What to expect when the unexpected happens: Social media communications across crises. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'15), pp. 994–1009 (2015)
76. Zubiaga, A., Liakata, M., Procter, R., Hoi, G.W.S., Tolmie, P.: Analysing how people orient to and spread rumours in social media by looking at conversational threads. PloS one **11**(3), 0150989 (2016)
77. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95), pp. 1137–1145 (1995)
78. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11), pp. 262–272 (2011)
79. Yao, L., Zhang, Y., Wei, B., Qian, H., Wang, Y.: Incorporating probabilistic knowledge into topic models. In: Proceedings of the 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'15), pp. 586–597 (2015)
80. Ritter, A., Etzioni, O., Clark, S.: Open domain event extraction from twitter. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12), pp. 1104–1112 (2012)
81. Halder, S., Lim, K.H., Chan, J., Zhang, X.: Transformer-based multi-task learning for queuing time aware next poi recommendation. In: Proceedings of the 25th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'21), pp. 510–523 (2021)
82. Brilhante, I.R., Macedo, J.A., Nardini, F.M., Perego, R., Renso, C.: On planning sightseeing tours with tripbuilder. Information Processing & Management **51**(2), 1–15 (2015)
83. Zhou, F., Wu, H., Trajcevski, G., Khokhar, A., Zhang, K.: Semi-supervised trajectory understanding with poi attention for end-to-end trip recommendation. ACM Transactions on Spatial Algorithms and Systems (TSAS) **6**(2), 1–25 (2020)
84. Zheng, D., Hu, T., You, Q., Kautz, H.A., Luo, J.: Towards lifestyle understanding: Predicting home and vacation locations from user's online photo collections. In: Proceedings of the Ninth International AAAI Conference on Web and Social Media (KDD'15), pp. 553–561 (2015)
85. Cao, B., Chen, F., Joshi, D., Philip, S.Y.: Inferring crowd-sourced venues for tweets. In: Proceedings of the 2015 IEEE International Conference on Big Data (BigData'15), pp. 639–648 (2015)
86. Zheng, X., Han, J., Sun, A.: A survey of location prediction on twitter. IEEE Transactions on Knowledge and Data Engineering **30**(9), 1652–1671 (2018)
87. Dhiman, A., Toshniwal, D.: An approximate model for event detection from twitter data. IEEE Access **8**, 122168–122184 (2020)
88. George, Y., Karunasekera, S., Harwood, A., Lim, K.H.: Real-time spatio-temporal event detection on geotagged social media. Journal of Big Data **8**(91), 1–28 (2021)

89. Weng, J., Lim, E.-P., Jiang, J., He, Q.: Twitterrank: Finding topic-sensitive influential twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM'10), pp. 261–270 (2010)

Appendix A: Detailed Results on Topic Coherence and Pointwise Mutual Information

Table 4 Comparison of ClusTop algorithm against various baselines, in terms of Topic Coherence (TC) and Pointwise Mutual Information (PMI) for the top 5 and 10 keywords. The rank of an algorithm's performance for each metric are provided in brackets.

	Top 5 Keywords / Unigrams							**Top 10 Keywords / Unigrams**						
	Dataset A		**Dataset B**		**Dataset C**		**Average**	**Dataset A**		**Dataset B**		**Dataset C**		**Average**
Algorithm	TC	PMI	TC	PMI	TC	PMI	**Rank@5**	TC	PMI	TC	PMI	TC	PMI	**Rank@10**
ClusTop-Word-NA	-37.6 (21)	-5.5 (15)	-34.1 (15)	-7.7 (14)	-37.9 (18)	-14.4 (17)	(16.7)	-171.0 (21)	-49.2 (16)	-160.8 (17)	-39.5 (15)	-173.4 (18)	-67.5 (19)	(17.7)
ClusTop-BiG-NA	-36.6 (20)	7.3 (8)	-35.9 (17)	1.2 (9)	-42.5 (22)	-16.4 (18)	(15.7)	-153.4 (20)	-29.6 (13)	-158.2 (16)	-25.8 (13)	-194.8 (21)	-63.4 (18)	(16.8)
ClusTop-TriG-NA	-30.9 (17)	10.7 (5)	-35.8 (16)	-2.6 (13)	-42.0 (21)	-18.2 (19)	(15.2)	-122.6 (16)	-16.1 (12)	-166.5 (19)	-25.1 (12)	-194.2 (20)	-73.5 (21)	(16.7)
ClusTop-BiHa-NA	-23.3 (12)	19.6 (1)	-32.3 (14)	4.7 (7)	-37.9 (18)	-11.2 (14)	(11.0)	-81.4 (12)	7.1 (4)	-140.8 (15)	-14.9 (11)	-169.9 (17)	-50.7 (15)	(12.3)
ClusTop-Hash-NA	-7.1 (2)	5.8 (9)	-14.8 (4)	0.3 (11)	-14.1 (4)	2.6 (6)	(6.0)	-19.4 (2)	2.3 (8)	-54.9 (5)	-6.9 (8)	-47.8 (4)	4.4 (5)	(5.3)
ClusTop-Noun-NA	-17.1 (6)	10.6 (6)	-21.4 (8)	6.9 (5)	-22.8 (10)	-0.3 (10)	(7.5)	-64.7 (8)	2.9 (6)	-90.5 (10)	-3.6 (7)	-97.8 (14)	-14.1 (10)	(9.2)
ClusTop-H2VG-NA	-17.9 (8)	-7.8 (16)	-22.8 (10)	-9.3 (15)	-22.5 (8)	-7.9 (12)	(11.5)	-69.3 (9)	-35.7 (14)	-87.2 (9)	-35.5 (14)	-62.1 (7)	-20.9 (11)	(10.7)
ClusTop-H2VW-NA	-9.4 (3)	16.6 (3)	-11.0 (2)	18.2 (1)	-8.9 (2)	21.2 (1)	(2.0)	-28.0 (3)	40.3 (2)	-32.5 (2)	48.4 (1)	-18.4 (1)	48.6 (2)	(1.8)
ClusTop-H2VF-NA	-10.5 (4)	18.8 (2)	-10.4 (1)	18.0 (2)	-8.6 (1)	20.3 (2)	(2.0)	-31.8 (5)	45.7 (1)	-30.7 (1)	45.5 (2)	-19.0 (2)	48.7 (1)	(2.0)
ClusTop-Word-AH	-30.9 (17)	-1.6 (14)	-40.2 (19)	-27.6 (19)	-24.2 (11)	10.3 (3)	(13.8)	-137.6 (18)	-57.7 (17)	-198.3 (21)	-131.1 (20)	-88.5 (12)	9.1 (4)	(15.3)
ClusTop-Hash-AH	-6.3 (1)	5.4 (10)	-12.8 (3)	0.9 (10)	-13.1 (3)	2.6 (6)	(5.5)	-16.2 (1)	1.5 (9)	-47.3 (3)	-7.3 (9)	-43.7 (3)	2.2 (6)	(5.2)
ClusTop-Noun-AH	-28.7 (14)	-11.4 (18)	-41.8 (20)	-19.8 (18)	-17.6 (5)	4.6 (4)	(13.2)	-132.4 (17)	-72.3 (19)	-185.6 (20)	-102.5 (18)	-63.8 (8)	-2.5 (9)	(15.2)
ClusTop-H2VG-AH	-17.6 (7)	-9.1 (17)	-32.0 (13)	-15.6 (17)	-29.5 (14)	-11.3 (15)	(13.8)	-71.2 (10)	-41.0 (15)	-136.3 (14)	-64.7 (17)	-97.9 (15)	-33.7 (14)	(14.2)
ClusTop-H2VW-AH	-27.2 (13)	-23.8 (20)	-38.7 (18)	-32.1 (20)	-26.6 (13)	-19.2 (21)	(17.5)	-84.0 (13)	-83.3 (20)	-133.6 (13)	-113.9 (19)	-87.7 (11)	-60.7 (16)	(15.3)
ClusTop-H2VF-AH	-29.0 (15)	-23.8 (21)	-45.1 (21)	-38.4 (21)	-25.7 (12)	-18.3 (20)	(18.3)	-97.3 (14)	-93.3 (21)	-166.1 (18)	-144.0 (21)	-85.1 (10)	-62.9 (17)	(16.8)
ClusTop-Word-AM	-34.4 (19)	8.3 (7)	-30.9 (12)	11.0 (4)	-37.7 (17)	-14.3 (16)	(12.5)	-146.1 (19)	-5.4 (11)	-126.8 (12)	8.1 (4)	-179.9 (19)	-69.0 (20)	(14.2)
ClusTop-Hash-AM	-19.7 (11)	11.4 (4)	-18.5 (5)	16.3 (3)	-33.7 (16)	-7.3 (11)	(8.3)	-73.9 (11)	8.5 (3)	-52.9 (4)	14.3 (3)	-153.6 (16)	-30.4 (12)	(8.2)
ClusTop-Noun-AM	-11.2 (5)	4.8 (12)	-19.2 (6)	0.3 (11)	-22.6 (9)	4.0 (5)	(8.0)	-29.9 (4)	-0.3 (10)	-70.1 (8)	-9.0 (10)	-70.3 (9)	11.9 (3)	(7.3)
ClusTop-H2VG-AM	-30.9 (17)	-13.9 (19)	-26.7 (11)	-12.0 (16)	-29.5 (14)	-10.9 (13)	(15.0)	-119.1 (15)	-59.3 (18)	-99.4 (11)	-43.6 (16)	-95.7 (13)	-31.9 (13)	(14.3)
ClusTop-H2VW-AM	-18.7 (9)	3.7 (13)	-21.6 (9)	3.1 (8)	-19.9 (7)	0.7 (9)	(9.2)	-53.3 (6)	2.7 (7)	-69.6 (7)	6.9 (5)	-58.8 (6)	0.3 (8)	(6.5)
ClusTop-H2VF-AM	-19.6 (10)	5.2 (11)	-20.2 (7)	5.0 (6)	-19.6 (6)	1.9 (8)	(8.0)	-54.0 (7)	3.4 (5)	-65.2 (6)	5.4 (6)	-58.2 (5)	0.4 (7)	(6.0)
LDA-Orig	-74.7 (24)	-74.4 (24)	-66.9 (24)	-62.2 (24)	-54.2 (24)	-43.1 (24)	(24.0)	-323.5 (24)	-307.5 (24)	-297.1 (24)	-269.2 (24)	-251.3 (24)	-191.9 (24)	(24.0)
LDA-Hash	-51.2 (22)	-43.8 (22)	-55.1 (23)	-42.9 (22)	-41.5 (20)	-23.4 (22)	(21.8)	-247.1 (22)	-185.4 (22)	-256.8 (22)	-199.2 (22)	-206.6 (22)	-112.5 (22)	(22.0)
LDA-Ment	-52.8 (23)	-45.9 (23)	-54.1 (22)	-45.3 (23)	-47.3 (23)	-27.7 (23)	(22.8)	-250.2 (23)	-198.8 (23)	-258.6 (23)	-206.5 (23)	-225.5 (23)	-136.4 (23)	(23.0)

Table 5 Comparison of ClusTop algorithm against various baselines, in terms of Topic Coherence (TC) and Pointwise Mutual Information (PMI) for the top 15 and 20 keywords. The rank of an algorithm's performance for each metric are provided in brackets.

	Top 15 Keywords / Unigrams							Top 20 Keywords / Unigrams						
	Dataset A		Dataset B		Dataset C		Average	Dataset A		Dataset B		Dataset C		Average
Algorithm	TC	PMI	TC	PMI	TC	PMI	**Rank@15**	TC	PMI	TC	PMI	TC	PMI	**Rank@20**
ClusTop-Word-NA	-409.7 (21)	-157.4 (17)	-382.6 (19)	-115.5 (16)	-403.5 (18)	-147.6 (21)	(18.7)	-768.2 (21)	-318.1 (19)	-695.1 (19)	-234.8 (17)	-732.7 (18)	-247.8 (21)	(19.2)
ClusTop-BiG-NA	-357.2 (20)	-108.3 (15)	-360.3 (17)	-77.2 (12)	-447.9 (20)	-116.5 (18)	(17.0)	-645.5 (20)	-212.3 (16)	-633.7 (17)	-147.8 (13)	-794.7 (20)	-165.4 (18)	(17.3)
ClusTop-TriG-NA	-289.2 (16)	-83.1 (13)	-382.0 (18)	-86.2 (15)	-451.6 (21)	-145.4 (20)	(17.2)	-530.2 (16)	-170.0 (14)	-674.8 (18)	-183.1 (14)	-804.0 (21)	-222.8 (19)	(17.0)
ClusTop-BiHa-NA	-179.9 (13)	-24.2 (10)	-324.4 (15)	-82.8 (14)	-389.5 (17)	-98.6 (17)	(14.3)	-319.3 (13)	-64.1 (10)	-569.2 (15)	-189.6 (15)	-692.0 (17)	-141.3 (17)	(14.5)
ClusTop-Hash-NA	-38.4 (2)	-4.0 (5)	-116.4 (5)	-11.3 (7)	-98.6 (6)	13.1 (4)	(4.8)	-63.7 (2)	-11.0 (6)	-195.2 (7)	-9.7 (6)	-148.8 (9)	28.2 (4)	(5.7)
ClusTop-Noun-NA	-139.1 (8)	-23.3 (9)	-208.5 (11)	-31.4 (9)	-219.9 (15)	-29.6 (11)	(10.5)	-242.1 (8)	-63.7 (9)	-363.5 (11)	-86.9 (9)	-384.7 (15)	-36.4 (10)	(10.3)
ClusTop-H2VG-NA	-160.2 (10)	-77.7 (12)	-183.4 (9)	-70.6 (11)	-75.6 (3)	-24.8 (10)	(9.2)	-292.1 (11)	-127.5 (11)	-295.8 (10)	-107.8 (11)	-75.6 (3)	-24.8 (8)	(9.0)
ClusTop-H2VW-NA	-49.9 (3)	72.7 (2)	-61.4 (2)	96.8 (1)	-19.3 (1)	52.5 (2)	(1.8)	-73.4 (3)	111.3 (2)	-95.7 (2)	161.4 (1)	-19.3 (1)	52.5 (3)	(2.0)
ClusTop-H2VF-NA	-59.5 (4)	87.5 (1)	-60.0 (1)	94.2 (2)	-20.7 (2)	54.8 (1)	(1.8)	-92.4 (4)	145.3 (1)	-95.0 (1)	157.7 (2)	-20.7 (2)	54.8 (2)	(2.0)
ClusTop-Word-AH	-311.9 (18)	-172.7 (19)	-480.2 (21)	-306.9 (21)	-177.0 (14)	-14.8 (8)	(16.8)	-567.7 (17)	-336.0 (21)	-903.7 (21)	-562.7 (21)	-298.8 (14)	-50.2 (13)	(17.8)
ClusTop-Hash-AH	-32.3 (1)	-4.2 (6)	-99.8 (3)	-16.3 (8)	-90.7 (5)	8.7 (5)	(4.7)	-54.8 (1)	-9.5 (5)	-169.0 (3)	-22.2 (7)	-143.6 (6)	21.9 (5)	(4.5)
ClusTop-Noun-AH	-309.9 (17)	-173.6 (20)	-444.0 (20)	-242.1 (19)	-131.8 (9)	-17.5 (9)	(15.7)	-575.1 (18)	-316.4 (18)	-811.4 (20)	-425.6 (19)	-221.7 (12)	-35.0 (9)	(16.0)
ClusTop-H2VG-AH	-166.4 (11)	-96.1 (14)	-304.6 (14)	-127.6 (17)	-143.0 (12)	-45.3 (13)	(13.5)	-303.3 (12)	-166.6 (13)	-530.9 (14)	-195.1 (16)	-147.6 (8)	-46.9 (12)	(12.5)
ClusTop-H2VW-AH	-167.4 (12)	-166.8 (18)	-281.9 (13)	-237.7 (18)	-116.9 (8)	-79.1 (15)	(14.0)	-274.1 (10)	-268.2 (17)	-476.9 (12)	-390.7 (18)	-122.6 (5)	-82.5 (15)	(12.8)
ClusTop-H2VF-AH	-201.1 (14)	-197.3 (21)	-350.3 (16)	-299.5 (20)	-134.5 (10)	-95.8 (16)	(16.2)	-335.0 (14)	-329.3 (20)	-591.2 (16)	-494.6 (20)	-160.1 (10)	-115.3 (16)	(16.0)
ClusTop-Word-AM	-326.6 (19)	-57.3 (11)	-278.7 (12)	-42.5 (10)	-422.5 (19)	-143.4 (19)	(15.0)	-599.9 (19)	-155.4 (12)	-492.7 (13)	-122.3 (12)	-766.5 (19)	-239.6 (20)	(15.8)
ClusTop-Hash-AM	-152.7 (9)	-17.5 (8)	-102.1 (4)	-1.8 (5)	-350.0 (16)	-50.1 (14)	(9.3)	-262.2 (9)	-56.7 (8)	-169.8 (4)	-26.5 (8)	-615.2 (16)	-67.0 (14)	(9.8)
ClusTop-Noun-AM	-60.7 (5)	-6.0 (7)	-150.6 (8)	-10.5 (6)	-138.9 (11)	29.7 (3)	(6.7)	-99.3 (5)	-11.4 (7)	-257.4 (8)	-2.4 (5)	-227.4 (13)	58.0 (1)	(6.5)
ClusTop-H2VG-AM	-273.6 (15)	-123.9 (16)	-193.0 (10)	-78.1 (13)	-152.4 (13)	-43.1 (12)	(13.2)	-480.9 (15)	-202.2 (15)	-289.8 (9)	-103.9 (10)	-167.3 (11)	-46.4 (11)	(11.8)
ClusTop-H2VW-AM	-95.4 (6)	4.1 (4)	-128.1 (7)	25.2 (3)	-89.3 (4)	5.5 (6)	(5.0)	-146.2 (6)	9.5 (4)	-192.2 (6)	53.3 (3)	-102.6 (4)	12.2 (7)	(5.0)
ClusTop-H2VF-AM	-96.2 (7)	4.3 (3)	-122.6 (6)	14.9 (4)	-108.3 (7)	4.1 (7)	(5.7)	-149.5 (7)	11.3 (3)	-185.6 (5)	30.2 (4)	-146.6 (7)	13.0 (6)	(5.3)
LDA-Orig	-722.5 (24)	-659.1 (24)	-695.0 (24)	-619.0 (24)	-593.5 (24)	-442.7 (24)	(24.0)	-1279.6 (24)	-1133.6 (24)	-1262.1 (24)	-1102.0 (24)	-1083.2 (24)	-790.0 (24)	(24.0)
LDA-Hash	-607.9 (22)	-435.4 (22)	-611.4 (22)	-449.2 (22)	-500.7 (22)	-271.5 (22)	(22.0)	-1134.2 (22)	-794.2 (22)	-1120.9 (22)	-805.6 (22)	-930.7 (22)	-463.1 (22)	(22.0)
LDA-Ment	-611.6 (23)	-465.7 (23)	-613.9 (23)	-477.7 (23)	-540.9 (23)	-316.9 (23)	(23.0)	-1141.5 (23)	-850.2 (23)	-1130.6 (23)	-855.9 (23)	-981.9 (23)	-566.6 (23)	(23.0)

Appendix B: Detailed Results on Precision, Recall and F-score

Table 6 Comparison of ClusTop algorithm against various baselines, in terms of Precision (Pre), Recall (Rec) and F-score (FS) for the top 5 keywords/unigrams of each topic. The rank of an algorithm's performance for each metric are provided in brackets.

					Top 5 Keywords / Unigrams					
		Dataset A			**Dataset B**			**Dataset C**		**Average**
Algorithm	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	**Rank@5**
ClusTop-Word-NA	.754±.002 (22)	.031±.000 (6)	.059±.000 (6)	.866±.003 (21)	.043±.000 (5)	.082±.001 (5)	.840±.009 (18)	.027±.001 (7)	.052±.001 (7)	(10.8)
ClusTop-BiG-NA	.786±.002 (18)	.034±.000 (2)	.064±.000 (2)	.857±.003 (22)	.046±.000 (2)	.086±.001 (2)	.833±.010 (20)	.029±.001 (4)	.056±.001 (3)	(8.3)
ClusTop-TriG-NA	.791±.002 (17)	.034±.000 (2)	.064±.000 (2)	.871±.003 (20)	.046±.000 (2)	.087±.001 (1)	.822±.009 (21)	.031±.001 (2)	.058±.001 (2)	(7.7)
ClusTop-BiHa-NA	.784±.002 (19)	.032±.000 (4)	.060±.000 (4)	.886±.003 (18)	.045±.000 (4)	.084±.001 (4)	.820±.009 (22)	.029±.001 (4)	.055±.001 (5)	(9.3)
ClusTop-Hash-NA	.898±.004 (10)	.023±.000 (21)	.044±.000 (18)	.916±.007 (13)	.032±.001 (17)	.062±.001 (17)	.936±.011 (8)	.022±.000 (17)	.042±.001 (18)	(15.4)
ClusTop-Noun-NA	.761±.002 (21)	.028±.000 (8)	.054±.000 (8)	.836±.003 (24)	.043±.000 (5)	.081±.001 (6)	.888±.008 (14)	.032±.001 (1)	.062±.001 (1)	(9.8)
ClusTop-H2VG-NA	.924±.003 (3)	.023±.000 (21)	.045±.000 (17)	.978±.004 (2)	.032±.001 (17)	.062±.001 (17)	.976±.007 (4)	.023±.001 (14)	.044±.001 (14)	(12.1)
ClusTop-H2VW-NA	.906±.001 (9)	.023±.001 (21)	.023±.001 (22)	.963±.001 (6)	.028±.001 (20)	.028±.001 (20)	.986±.001 (1)	.019±.001 (23)	.019±.001 (23)	(16.1)
ClusTop-H2VF-NA	.910±.001 (7)	.025±.001 (15)	.025±.001 (19)	.960±.001 (7)	.029±.001 (19)	.029±.001 (19)	.955±.001 (6)	.019±.001 (23)	.019±.001 (23)	(15.3)
ClusTop-Word-AH	.741±.003 (24)	.025±.000 (15)	.049±.000 (13)	.845±.005 (23)	.040±.000 (9)	.075±.001 (9)	.844±.010 (17)	.027±.001 (7)	.052±.001 (7)	(13.8)
ClusTop-Hash-AH	.847±.002 (12)	.026±.000 (12)	.051±.000 (12)	.912±.004 (15)	.046±.001 (2)	.086±.001 (2)	.896±.008 (12)	.024±.000 (12)	.046±.001 (13)	(10.2)
ClusTop-Noun-AH	.802±.002 (16)	.024±.000 (18)	.046±.000 (16)	.873±.004 (19)	.037±.000 (13)	.071±.001 (13)	.872±.008 (15)	.029±.001 (4)	.056±.001 (3)	(13.0)
ClusTop-H2VG-AH	.919±.002 (5)	.025±.000 (15)	.049±.000 (13)	.902±.003 (16)	.042±.000 (7)	.080±.001 (7)	.891±.008 (13)	.025±.001 (10)	.048±.001 (11)	(10.8)
ClusTop-H2VW-AH	.927±.001 (1)	.023±.001 (21)	.023±.001 (22)	.972±.001 (4)	.028±.001 (20)	.028±.001 (20)	.979±.001 (3)	.020±.001 (21)	.020±.001 (21)	(14.8)
ClusTop-H2VF-AH	.918±.001 (6)	.025±.001 (15)	.025±.001 (19)	.965±.001 (5)	.027±.001 (23)	.027±.001 (23)	.948±.001 (7)	.021±.001 (19)	.021±.001 (19)	(15.1)
ClusTop-Word-AM	.748±.001 (23)	.034±.000 (2)	.065±.000 (1)	.929±.003 (11)	.036±.000 (15)	.069±.001 (15)	.758±.016 (24)	.022±.000 (17)	.043±.001 (16)	(13.8)
ClusTop-Hash-AM	.763±.002 (20)	.027±.000 (10)	.052±.000 (10)	.917±.003 (12)	.037±.000 (13)	.072±.001 (12)	.869±.011 (16)	.024±.000 (12)	.047±.001 (12)	(13.0)
ClusTop-Noun-AM	.842±.002 (13)	.025±.000 (15)	.048±.000 (15)	.950±.003 (9)	.039±.000 (10)	.074±.001 (10)	.923±.009 (9)	.022±.000 (17)	.043±.001 (16)	(12.7)
ClusTop-H2VG-AM	.864±.002 (11)	.028±.000 (8)	.054±.000 (8)	.930±.003 (10)	.041±.000 (8)	.078±.001 (8)	.900±.008 (10)	.025±.000 (10)	.049±.001 (9)	(9.1)
ClusTop-H2VW-AM	.924±.001 (3)	.023±.001 (21)	.023±.001 (22)	.976±.001 (3)	.027±.001 (23)	.027±.001 (23)	.981±.001 (2)	.020±.001 (21)	.020±.001 (21)	(15.4)
ClusTop-H2VF-AM	.910±.001 (7)	.022±.001 (24)	.022±.001 (24)	.985±.001 (1)	.027±.001 (23)	.027±.001 (23)	.971±.001 (5)	.020±.001 (21)	.020±.001 (21)	(16.6)
LDA-Orig	.925±.001 (2)	.027±.000 (10)	.052±.000 (10)	.956±.002 (8)	.037±.000 (13)	.070±.001 (14)	.898±.010 (11)	.025±.000 (10)	.049±.001 (9)	(9.7)
LDA-Hash	.821±.002 (15)	.031±.000 (6)	.059±.000 (6)	.916±.003 (13)	.036±.000 (15)	.069±.001 (15)	.837±.010 (19)	.028±.001 (6)	.054±.001 (6)	(11.2)
LDA-Ment	.830±.002 (14)	.031±.000 (6)	.060±.000 (4)	.900±.003 (17)	.039±.000 (10)	.074±.001 (10)	.814±.018 (23)	.023±.000 (14)	.044±.001 (14)	(12.4)

Table 7 Comparison of ClusTop algorithm against various baselines, in terms of Precision (Pre), Recall (Rec) and F-score (FS) for the top 10 keywords/unigrams of each topic. The rank of an algorithm's performance for each metric are provided in brackets.

	Top 10 Keywords / Unigrams									
		Dataset A			Dataset B			Dataset C		Average
Algorithm	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	**Rank@10**
ClusTop-Word-NA	.690±.002 (24)	.033±.000 (7)	.062±.000 (7)	.804±.003 (20)	.051±.000 (3)	.095±.001 (3)	.764±.010 (22)	.033±.001 (6)	.062±.001 (6)	(10.9)
ClusTop-BiG-NA	.707±.002 (22)	.035±.000 (2)	.065±.000 (2)	.789±.004 (22)	.052±.000 (2)	.096±.001 (2)	.765±.010 (20)	.035±.001 (3)	.067±.002 (2)	(8.6)
ClusTop-TriG-NA	.717±.002 (21)	.034±.000 (4)	.064±.000 (4)	.811±.003 (19)	.053±.000 (1)	.098±.001 (1)	.746±.010 (23)	.036±.001 (1)	.068±.001 (1)	(8.3)
ClusTop-BiHa-NA	.719±.002 (20)	.034±.000 (4)	.064±.000 (4)	.782±.004 (24)	.051±.000 (3)	.095±.001 (3)	.765±.010 (20)	.035±.001 (3)	.065±.001 (3)	(9.3)
ClusTop-Hash-NA	.860±.004 (8)	.023±.000 (24)	.045±.000 (17)	.896±.008 (9)	.032±.001 (17)	.061±.001 (17)	.925±.012 (8)	.022±.000 (18)	.043±.001 (18)	(15.1)
ClusTop-Noun-NA	.736±.002 (16)	.033±.000 (7)	.062±.000 (7)	.803±.003 (21)	.049±.000 (5)	.091±.001 (5)	.802±.009 (16)	.035±.001 (3)	.065±.001 (3)	(9.2)
ClusTop-H2VG-NA	.904±.004 (3)	.024±.000 (21)	.046±.000 (16)	.973±.004 (1)	.032±.001 (17)	.061±.001 (17)	.964±.008 (3)	.023±.001 (16)	.044±.001 (16)	(12.2)
ClusTop-H2VW-NA	.894±.001 (5)	.026±.001 (17)	.026±.001 (22)	.944±.001 (4)	.029±.001 (20)	.029±.001 (20)	.983±.001 (1)	.020±.001 (23)	.020±.001 (23)	(15.0)
ClusTop-H2VF-NA	.892±.001 (6)	.026±.001 (17)	.026±.001 (22)	.931±.001 (8)	.029±.001 (20)	.029±.001 (20)	.943±.001 (7)	.021±.001 (20)	.021±.001 (20)	(15.6)
ClusTop-Word-AH	.692±.003 (23)	.025±.000 (19)	.049±.000 (14)	.784±.004 (23)	.043±.000 (8)	.080±.001 (8)	.778±.010 (18)	.030±.001 (8)	.057±.001 (8)	(14.3)
ClusTop-Hash-AH	.833±.002 (11)	.027±.000 (12)	.051±.000 (13)	.842±.004 (15)	.041±.000 (11)	.078±.001 (11)	.857±.009 (12)	.025±.000 (13)	.048±.001 (13)	(12.3)
ClusTop-Noun-AH	.735±.002 (17)	.024±.000 (21)	.045±.000 (17)	.827±.004 (17)	.041±.000 (11)	.076±.001 (14)	.828±.009 (13)	.031±.001 (7)	.059±.001 (7)	(13.8)
ClusTop-H2VG-AH	.824±.002 (12)	.027±.000 (12)	.053±.000 (12)	.850±.003 (12)	.043±.000 (8)	.080±.001 (8)	.869±.009 (11)	.028±.001 (9)	.053±.001 (9)	(10.3)
ClusTop-H2VW-AH	.908±.001 (1)	.024±.001 (21)	.024±.001 (24)	.944±.001 (4)	.027±.001 (23)	.027±.001 (23)	.964±.001 (4)	.020±.001 (23)	.020±.001 (23)	(16.2)
ClusTop-H2VF-AH	.905±.001 (2)	.027±.001 (15)	.027±.001 (20)	.934±.001 (6)	.029±.001 (20)	.029±.001 (20)	.949±.001 (5)	.021±.001 (20)	.021±.001 (20)	(14.2)
ClusTop-Word-AM	.734±.001 (18)	.036±.000 (1)	.068±.000 (1)	.828±.004 (16)	.039±.000 (16)	.073±.001 (16)	.709±.013 (24)	.023±.000 (16)	.044±.001 (16)	(13.8)
ClusTop-Hash-AM	.731±.002 (19)	.030±.000 (10)	.058±.000 (10)	.845±.004 (14)	.043±.000 (8)	.081±.001 (7)	.823±.011 (14)	.028±.001 (9)	.053±.001 (9)	(11.1)
ClusTop-Noun-AM	.860±.001 (8)	.024±.000 (21)	.047±.000 (15)	.932±.003 (7)	.040±.000 (14)	.076±.001 (14)	.908±.008 (9)	.024±.000 (14)	.047±.001 (14)	(12.9)
ClusTop-H2VG-AM	.800±.002 (13)	.031±.000 (9)	.060±.000 (9)	.875±.003 (11)	.041±.000 (11)	.078±.001 (11)	.882±.008 (10)	.027±.000 (11)	.051±.001 (11)	(10.7)
ClusTop-H2VW-AM	.904±.001 (3)	.027±.001 (15)	.027±.001 (20)	.964±.001 (2)	.028±.001 (22)	.028±.001 (22)	.972±.001 (2)	.021±.001 (20)	.021±.001 (20)	(14.0)
ClusTop-H2VF-AM	.891±.001 (7)	.027±.001 (15)	.027±.001 (20)	.951±.001 (3)	.027±.001 (23)	.027±.001 (23)	.944±.001 (6)	.020±.001 (23)	.020±.001 (23)	(15.9)
LDA-Orig	.848±.002 (10)	.029±.000 (11)	.056±.000 (11)	.885±.003 (10)	.040±.000 (14)	.076±.001 (14)	.808±.011 (15)	.026±.001 (12)	.051±.001 (11)	(12.0)
LDA-Hash	.759±.002 (14)	.034±.000 (4)	.064±.000 (4)	.847±.003 (13)	.041±.000 (11)	.078±.001 (11)	.778±.010 (18)	.034±.001 (5)	.064±.001 (5)	(9.4)
LDA-Ment	.752±.002 (15)	.033±.000 (7)	.063±.000 (6)	.820±.004 (18)	.044±.000 (6)	.082±.001 (6)	.787±.013 (17)	.024±.000 (14)	.047±.001 (14)	(11.4)

Table 8 Comparison of ClusTop algorithm against various baselines, in terms of Precision (Pre), Recall (Rec) and F-score (FS) for the top 15 keywords/unigrams of each topic. The rank of an algorithm's performance for each metric are provided in brackets.

	Top 15 Keywords / Unigrams									
	Dataset A			Dataset B			Dataset C			Average
Algorithm	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	**Rank@15**
ClusTop-Word-NA	.648±.002 (24)	.034±.000 (5)	.063±.000 (7)	.743±.004 (23)	.053±.000 (3)	.098±.001 (3)	.717±.010 (22)	.038±.001 (4)	.070±.002 (4)	(10.6)
ClusTop-BiG-NA	.671±.002 (21)	.035±.000 (2)	.065±.000 (2)	.747±.004 (21)	.055±.001 (1)	.100±.001 (1)	.719±.010 (21)	.040±.001 (2)	.073±.002 (2)	(8.1)
ClusTop-TriG-NA	.667±.002 (22)	.035±.000 (2)	.065±.000 (2)	.757±.004 (20)	.055±.000 (1)	.100±.001 (1)	.677±.010 (24)	.041±.001 (1)	.075±.002 (1)	(8.2)
ClusTop-BiHa-NA	.667±.002 (22)	.034±.000 (5)	.064±.000 (5)	.729±.004 (24)	.052±.000 (4)	.096±.001 (4)	.722±.009 (20)	.040±.001 (2)	.073±.002 (2)	(9.8)
ClusTop-Hash-NA	.848±.004 (9)	.023±.000 (24)	.045±.000 (18)	.880±.008 (9)	.031±.001 (17)	.060±.001 (17)	.914±.012 (8)	.023±.001 (17)	.044±.001 (17)	(15.1)
ClusTop-Noun-NA	.698±.002 (18)	.033±.000 (8)	.061±.000 (9)	.773±.003 (19)	.051±.000 (5)	.093±.001 (5)	.753±.010 (15)	.036±.001 (6)	.067±.001 (6)	(10.1)
ClusTop-H2VG-NA	.895±.004 (2)	.024±.000 (23)	.046±.000 (17)	.961±.005 (1)	.031±.001 (17)	.060±.001 (17)	.962±.008 (3)	.023±.001 (17)	.044±.001 (17)	(12.7)
ClusTop-H2VW-NA	.889±.001 (5)	.026±.001 (21)	.026±.001 (24)	.935±.001 (3)	.030±.001 (20)	.030±.001 (20)	.983±.001 (1)	.020±.001 (23)	.020±.001 (23)	(15.6)
ClusTop-H2VF-NA	.862±.001 (7)	.028±.001 (14)	.028±.001 (19)	.903±.001 (8)	.030±.001 (20)	.030±.001 (20)	.942±.001 (6)	.021±.001 (20)	.021±.001 (20)	(14.9)
ClusTop-Word-AH	.677±.002 (20)	.025±.000 (22)	.048±.000 (16)	.744±.004 (22)	.043±.000 (12)	.079±.001 (14)	.731±.010 (19)	.033±.001 (7)	.061±.001 (8)	(15.6)
ClusTop-Hash-AH	.802±.002 (10)	.030±.000 (12)	.058±.000 (12)	.831±.004 (13)	.044±.000 (10)	.083±.001 (10)	.847±.009 (12)	.026±.001 (15)	.050±.001 (15)	(12.1)
ClusTop-Noun-AH	.699±.002 (17)	.027±.000 (16)	.051±.000 (14)	.791±.004 (16)	.042±.000 (14)	.078±.001 (15)	.781±.010 (13)	.033±.001 (7)	.062±.001 (7)	(13.2)
ClusTop-H2VG-AH	.762±.002 (12)	.029±.000 (13)	.055±.000 (13)	.832±.003 (12)	.043±.000 (12)	.080±.001 (12)	.859±.009 (11)	.028±.001 (10)	.053±.001 (11)	(11.8)
ClusTop-H2VW-AH	.902±.001 (1)	.027±.001 (18)	.027±.001 (22)	.913±.001 (5)	.029±.001 (22)	.029±.001 (22)	.964±.001 (2)	.020±.001 (23)	.020±.001 (23)	(15.3)
ClusTop-H2VF-AH	.893±.001 (3)	.028±.001 (14)	.028±.001 (19)	.913±.001 (5)	.030±.001 (20)	.030±.001 (20)	.945±.001 (5)	.021±.001 (20)	.021±.001 (20)	(14.0)
ClusTop-Word-AM	.687±.001 (19)	.036±.000 (1)	.067±.000 (1)	.790±.004 (17)	.046±.000 (7)	.085±.001 (7)	.689±.012 (23)	.027±.001 (13)	.052±.001 (13)	(11.2)
ClusTop-Hash-AM	.713±.002 (15)	.033±.000 (8)	.063±.000 (7)	.816±.004 (14)	.045±.000 (9)	.083±.001 (10)	.770±.011 (14)	.030±.001 (9)	.058±.001 (9)	(10.6)
ClusTop-Noun-AM	.860±.001 (8)	.026±.000 (20)	.050±.000 (15)	.905±.003 (7)	.041±.000 (16)	.077±.001 (16)	.892±.008 (9)	.026±.000 (15)	.049±.001 (16)	(13.6)
ClusTop-H2VG-AM	.758±.002 (13)	.031±.000 (11)	.060±.000 (10)	.863±.003 (10)	.044±.000 (10)	.083±.001 (10)	.876±.008 (10)	.027±.000 (13)	.052±.001 (13)	(11.1)
ClusTop-H2VW-AM	.892±.001 (4)	.027±.001 (18)	.027±.001 (22)	.942±.001 (2)	.028±.001 (23)	.028±.001 (23)	.960±.001 (4)	.021±.001 (20)	.021±.001 (20)	(15.1)
ClusTop-H2VF-AM	.877±.001 (6)	.027±.001 (18)	.027±.001 (22)	.935±.001 (3)	.027±.001 (24)	.027±.001 (24)	.931±.001 (7)	.021±.001 (20)	.021±.001 (20)	(16.0)
LDA-Orig	.799±.002 (11)	.032±.000 (10)	.060±.000 (10)	.840±.003 (11)	.042±.000 (14)	.080±.001 (12)	.753±.010 (15)	.028±.001 (10)	.054±.001 (10)	(11.4)
LDA-Hash	.720±.002 (14)	.034±.000 (5)	.064±.000 (5)	.795±.003 (15)	.046±.000 (7)	.085±.001 (7)	.733±.010 (18)	.038±.001 (4)	.070±.001 (4)	(8.8)
LDA-Ment	.703±.002 (16)	.034±.000 (5)	.064±.000 (5)	.774±.004 (18)	.048±.000 (6)	.088±.001 (6)	.743±.012 (17)	.027±.001 (13)	.052±.001 (13)	(11.0)

Table 9 Comparison of ClusTop algorithm against various baselines, in terms of Precision (Pre), Recall (Rec) and F-score (FS) for the top 20 keywords/unigrams of each topic. The rank of an algorithm's performance for each metric are provided in brackets.

				Top 20 Keywords / Unigrams						
		Dataset A			**Dataset B**			**Dataset C**		**Average**
Algorithm	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	**Rank@20**
ClusTop-Word-NA	.617±.002 (24)	.034±.000 (6)	.064±.000 (6)	.709±.003 (23)	.055±.000 (2)	.100±.001 (3)	.669±.009 (22)	.038±.001 (4)	.070±.002 (4)	(10.4)
ClusTop-BiG-NA	.646±.002 (21)	.035±.000 (3)	.065±.000 (3)	.722±.004 (22)	.056±.001 (1)	.102±.001 (1)	.679±.010 (20)	.040±.001 (2)	.073±.002 (2)	(8.3)
ClusTop-TriG-NA	.636±.002 (22)	.035±.000 (3)	.065±.000 (3)	.726±.004 (20)	.055±.001 (2)	.101±.001 (2)	.662±.009 (24)	.041±.001 (1)	.075±.002 (1)	(8.7)
ClusTop-BiHa-NA	.635±.002 (23)	.035±.000 (3)	.065±.000 (3)	.702±.004 (24)	.054±.000 (4)	.099±.001 (4)	.679±.009 (20)	.040±.001 (2)	.073±.002 (2)	(9.4)
ClusTop-Hash-NA	.831±.004 (9)	.024±.000 (23)	.045±.000 (18)	.873±.008 (9)	.031±.001 (17)	.060±.001 (17)	.911±.012 (8)	.023±.001 (17)	.044±.001 (17)	(15.0)
ClusTop-Noun-NA	.666±.002 (17)	.033±.000 (9)	.061±.000 (11)	.748±.003 (18)	.051±.000 (5)	.094±.001 (5)	.719±.010 (15)	.036±.001 (6)	.067±.001 (6)	(10.2)
ClusTop-H2VG-NA	.891±.004 (2)	.024±.000 (23)	.046±.000 (17)	.952±.005 (1)	.031±.001 (17)	.060±.001 (17)	.962±.008 (2)	.023±.001 (17)	.044±.001 (17)	(12.6)
ClusTop-H2VW-NA	.869±.001 (5)	.027±.001 (20)	.027±.001 (22)	.927±.001 (4)	.030±.001 (20)	.030±.001 (20)	.983±.001 (1)	.020±.001 (23)	.020±.001 (23)	(15.3)
ClusTop-H2VF-NA	.856±.001 (7)	.028±.001 (16)	.028±.001 (19)	.894±.001 (7)	.030±.001 (20)	.030±.001 (20)	.942±.001 (5)	.021±.001 (20)	.021±.001 (20)	(14.9)
ClusTop-Word-AH	.660±.002 (19)	.029±.000 (14)	.055±.000 (14)	.724±.004 (21)	.044±.001 (12)	.080±.001 (13)	.703±.010 (17)	.033±.001 (7)	.061±.001 (8)	(13.9)
ClusTop-Hash-AH	.757±.002 (10)	.033±.000 (9)	.063±.000 (8)	.801±.004 (13)	.044±.000 (12)	.082±.001 (11)	.836±.009 (12)	.026±.001 (15)	.050±.001 (15)	(11.7)
ClusTop-Noun-AH	.700±.002 (14)	.029±.000 (14)	.054±.000 (15)	.774±.004 (16)	.042±.000 (15)	.079±.001 (15)	.743±.010 (14)	.033±.001 (7)	.062±.001 (7)	(13.0)
ClusTop-H2VG-AH	.743±.002 (12)	.030±.000 (13)	.058±.000 (13)	.809±.003 (11)	.043±.000 (14)	.080±.001 (13)	.858±.009 (11)	.028±.001 (10)	.053±.001 (11)	(12.0)
ClusTop-H2VW-AH	.894±.001 (1)	.027±.001 (20)	.027±.001 (22)	.902±.001 (6)	.029±.001 (22)	.029±.001 (22)	.962±.001 (3)	.020±.001 (23)	.020±.001 (23)	(15.8)
ClusTop-H2VF-AH	.879±.001 (3)	.028±.001 (16)	.028±.001 (19)	.905±.001 (5)	.030±.001 (20)	.030±.001 (20)	.939±.001 (6)	.021±.001 (20)	.021±.001 (20)	(14.3)
ClusTop-Word-AM	.660±.001 (19)	.036±.000 (1)	.067±.000 (1)	.778±.004 (15)	.050±.000 (6)	.091±.001 (6)	.664±.012 (23)	.027±.001 (13)	.052±.001 (13)	(10.8)
ClusTop-Hash-AM	.691±.002 (15)	.033±.000 (9)	.062±.000 (9)	.796±.004 (14)	.047±.000 (9)	.087±.001 (9)	.745±.011 (13)	.030±.001 (9)	.058±.001 (9)	(10.7)
ClusTop-Noun-AM	.841±.001 (8)	.027±.000 (18)	.051±.000 (16)	.883±.003 (8)	.041±.000 (16)	.078±.001 (16)	.859±.009 (10)	.026±.000 (15)	.049±.001 (16)	(13.7)
ClusTop-H2VG-AM	.736±.002 (13)	.032±.000 (11)	.061±.000 (11)	.856±.003 (10)	.045±.000 (10)	.084±.001 (10)	.875±.008 (9)	.027±.000 (13)	.052±.001 (13)	(11.1)
ClusTop-H2VW-AM	.876±.001 (4)	.027±.001 (20)	.027±.001 (22)	.936±.001 (2)	.029±.001 (22)	.029±.001 (22)	.955±.001 (4)	.021±.001 (20)	.021±.001 (20)	(15.1)
ClusTop-H2VF-AM	.859±.001 (6)	.027±.001 (20)	.027±.001 (22)	.930±.001 (3)	.028±.001 (24)	.028±.001 (24)	.928±.001 (7)	.021±.001 (20)	.021±.001 (20)	(16.2)
LDA-Orig	.752±.002 (11)	.032±.000 (11)	.061±.000 (11)	.802±.003 (12)	.044±.000 (12)	.082±.001 (11)	.706±.010 (16)	.028±.001 (10)	.054±.001 (10)	(11.6)
LDA-Hash	.689±.002 (16)	.035±.000 (3)	.065±.000 (3)	.751±.003 (17)	.048±.000 (8)	.088±.001 (8)	.702±.010 (18)	.038±.001 (4)	.070±.001 (4)	(9.0)
LDA-Ment	.666±.002 (17)	.034±.000 (6)	.064±.000 (6)	.730±.004 (19)	.049±.000 (7)	.090±.001 (7)	.698±.012 (19)	.027±.001 (13)	.052±.001 (13)	(11.9)