Social Network Analysis and Mining (2022) 12:89
https://doi.org/10.1007/s13278-022-00898-5

ORIGINAL ARTICLE

An effective short-text topic modelling with neighbourhood assistance-driven NMF in Twitter

Shalani Athukorala[1] · **Wathsala Mohotti**[1]

Received: 28 November 2021 / Revised: 29 April 2022 / Accepted: 30 April 2022

Abstract
Social media such as Twitter connect billions of people by allowing them to exchange their thoughts via short-text communication. Topic modelling is a widely used technique for analysing short texts. Discovering topic clusters in short-text collections faces issues with distance-based, density-based and dimensionality reduction-based methods due to their higher dimensionality and short length which results in extremely sparse text representation matrices. We propose the 'neighbourhood-based assistance'-driven non-negative matrix factorization (NMF) method to handle high-dimensional sparse short-text representation with lower-dimensional projection effectively. We utilized NMF that aligned with the natural non-negativity of text data coupled with the symmetric document affinity information to identify topic distribution in the short text. Neighbourhood information within documents is captured using Jaccard similarity to assist information loss, resulting in higher-to-lower-dimensional projection. Experimental results with Twitter data sets show that the proposed approach is able to attain high accuracy compared to state-of-the-art methods quantitatively, while qualitative analysis with case studies validates the ability of the proposed approach in generating meaningful topic clusters.

Keywords Short texts · Twitter · Topic modelling · Text mining · Non-negative matrix factorization · Neighbourhood

1 Introduction

The growth of text data is exponential over the Internet, and it is predicted to gain significant growth and attention in the coming years (Muthu et al. 2021). The popularity of the Internet, web platforms such as forums, Wikis, blogs, news feeds, e-marketplaces as well as emails play an important role in facilitating the exchange of opinions and social views through users' text communication. Among them, social media platforms have become extremely popular in disseminating trending information based on short-text communication over past years (Salloum et al. 2017). With the COVID-19 pandemic, users commonly rely on platforms like Twitter for social connectivity and news (Balasubramaniam et al. 2021). As the short-text data kept increasing, researchers and a broader audience of text analysis practitioners were interested in learning interesting patterns from them with different text mining methods. A topic model is an effective tool for information retrieval in social media analysis for event tracking, community discovery and text classification (Salloum et al. 2017).

Topic modelling follows an unsupervised approach that outputs topic clusters in unlabelled text with their term associations, thus resulting in better interpretation. A topic model discovers documents that share the same topic structure that occur in a document collection by finding the correlations of words that appear in documents (Blei 2012). Although there exists a myriad of topic modelling techniques that work well for document collections, they are impaired in handling short texts due to various reasons (Mohotti 2020). Usually, text matrix representation with a higher number of dimensions results in a sparse nature that challenges many text mining methods including topic modelling (Indah et al. 2019; Shi et al. 2018). The distance differences between points become negligible due to sparseness in high dimensions (Köppen 2000). This phenomenon known as distance concentration does not allow differentiation among documents based on distance differences.

✉ Shalani Athukorala
athukorala9910@usci.ruh.ac.lk

Wathsala Mohotti
wathsala@dcs.ruh.ac.lk

[1] Department of Computer Science, University of Ruhuna, Wellamadama, Matara 81000, Sri Lanka

Published online: 24 July 2022

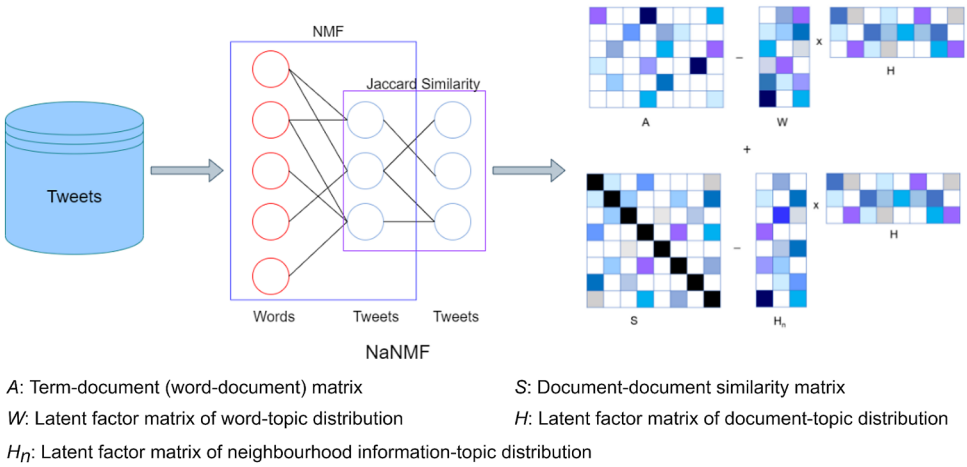


Fig. 1 Overview of the proposed NaNMF algorithm

The short text is a special case where extreme sparse text representation results from its comparatively shorter document length. This escalated the issue of distance concentration in distance-based and density estimation in density-based text mining methods (Mohotti and Nayak 2018). Researchers have developed probabilistic (Blei et al. 2003; Yan et al. 2013) and matrix factorization-based methods (Shi et al. 2018; Xu et al. 2003) to approximate the topic clusters in lower-dimensional space. However, the short length that results in low word co-occurrences challenges probability calculations, while matrix factorization-based methods result in information loss in lower-dimensional projection.

The non-negative matrix factorization (NMF) (Lee and Sebastian 1999) maps higher-dimensional data into lower-dimensional space providing solutions to issues in high-dimensionality. Given a non-negative matrix A, NMF finds a low-rank approximation to A as a multiplication of two non-negative matrices W and H (i.e. $A \approx WH$). This concept easily maps with the natural non-negativity in text data. Therefore, NMF-based methods are known to be effective in mining short texts (Mohotti 2020). Usually, NMF-based topic modelling approximates term–document matrix, A which is higher in dimensions, with lower rank basis vectors W that contain the word frequencies related to each topic, and H represents the associated weights for the topics in each document given the number of topics as lower rank k (Lee and Sebastian 1999; Yan et al. 2012). Thereby, matrix H is used to determine the topic of each document.

However, NMF results in an information loss within higher- to lower-dimensional projection. The information included in the sparse data is insufficient for effective lower-order approximation. Thus, the aforementioned information loss incurred significantly affects the performance in the case of highly sparse data such as short texts (Takeuchi et al. 2013). Therefore, researchers assist the matrix factorization process with additional information to compensate for this loss (Huang et al. 2020; Shi et al. 2018).

Assisting a topic model has been employed to retrieve more accurate results (Shi et al. 2018). In Huang et al. (2020), semantic assistance is used to improve the Biterm topic model which is a generative algorithm based on Gibbs sampling. A novel topic model called semantics-assisted non-negative matrix factorization (SeaNMF) that utilizes word–context semantic correlation information is developed to assist with the matrix factorization process (Shi et al. 2018). In Yangyang and Yin (2017), also an assistant-based topic model named latent embedding structured lifelong learning topic model (LLT model) is introduced which exploited pre-trained latent word embeddings and topic embeddings to provide assistance to a probabilistic topic model. However, semantic assistance based on simple term co-occurrences is not effective for short text with fewer terms, while word embedding methods are computationally complex with higher-dimensional data.

Motivated by the fact that assistance helps to improve topic models, we conjecture providing neighbourhood-based assistance for term–document matrix factorization improves the accuracy of short text topic models. It compensates for information loss in higher-to-lower-dimensional projection. More specifically, this paper proposes a novel neighbourhood assistance-based non-negative matrix factorization (NaNMF) that works in an unsupervised setting for short-text topic modelling. To the best of our knowledge, NaNMF is the first such method that coupled the Jaccard similarity-based document affinity matrix with the term–document matrix within the non-negative matrix factorization framework as shown in Fig. 1. Empirical analyses with several Twitter corpus reveal that NaNMF is able to identify topic clusters accurately compared to other state-of-the-art topic modelling methods.

The rest of the paper is organized as follows. Section 2 details the related work on topic modelling. The proposed approach and implementation are elaborated in Sect. 3. A comprehensive empirical study and benchmarks on several

Twitter data sets with relevant topic modelling algorithms are provided in Sect. 4. The final concluding remarks are presented in Sect. 5.

2 Literature review

Eighty per cent of the information available on the world wide web is currently stored in unstructured textual format (Sheikh 2017). News articles, content in social media, blog posts, customer reviews on e-commerce platforms, etc., fall into the unstructured text category. Also, billions of short texts are produced every day, in the form of search queries, ad keywords, tweets and other social media posts, etc. (Wang et al. 2016). This vast amount of digital text is a gold mine and can be utilized for several purposes when mined precisely.

Unusually, text mining methods work with text similarity calculation to identify the clusters (Muflikhah and Baharudin 2009; Ferdous 2009; Wu et al. 2020). Text similarity can either be lexical or semantic. Texts are lexically similar if they have a similar word sequence or structure while semantically similar if the words have the same theme or meaning (Gomaa and Fahmy 2013). Cosine similarity is a widely used lexical similarity measure. In there, the documents are represented as term vectors and the similarity corresponds to the correlation (cosine angle) between the vectors (Huang 2008). Jaccard is another lexical similarity measure that computed similarity as the number of shared terms over the number of all unique terms in both texts (Huang 2008). Sigmoid Jaccard's similarity is a recent, improved version of the common Jaccard similarity (Likavec et al. 2019). It is a feature-based formula developed by combining Jaccard similarity with the sigmoid function. In contrast, Word2Vec determines the semantic distance between words. In there, a network is trained by using a large external corpus and giving positive feedback when words appear together in context and giving negative feedback when words are randomly swapped into other contexts (Handler 2014). These similarity measures can be used to indicate the neighbourhood of a document in a document matrix representation.

The most important attribute of text data is that they are sparse and high dimensional (Aggarwal and Zhai 2012) which causes immediate implications for the aforementioned text similarity calculation measures and thereby the mining techniques. Additionally, short texts have some unique characteristics that distinguish them from normal documents. (1) Short texts contain a limited number of words. For example, the majority of search queries contain less than 5 words, and tweets can have no more than 280 characters. (2) Short texts usually do not follow the formal syntax of a written language, making them ambiguous and difficult to interpret. Due to the limited length of short text, the document representation matrix tends to be extremely sparse, making it inefficient for distance-based or density-based computations; the difference between data points becomes negligible and leads to distance concentration (Balbi 2010). Also, knowledge acquisition, representation and inference from short text require additional effort compared to other documents due to their unstructured phrases and symbols.

Topic modelling is one of the most powerful techniques in text mining for data mining. It performs latent topic discovery by finding relationships among terms, topics and text documents (Jelodar et al. 2019). Specifically, it finds the structure of words appearing in documents and how to link documents that share the same structure. Usually, unsupervised topic models closely match the needs of real-world applications; thus, it saves the effort of creating labelled data and training classifiers. Thus, topic modelling is used for various text mining tasks such as summarization, sentiment analysis and document classification (Jelodar et al. 2019).

Generally, topic modelling methods are classified as probabilistic and non-probabilistic in the literature (Kherwa and Bansal 2020). This section reviews and classifies the literature related to four main branches used to develop topic modelling algorithms: (1) traditional text mining methods for topic modelling, (2) probabilistic topic models, (3) neural topic models and (4) matrix factorization-based topic models.

2.1 Traditional text mining for topic modelling

K-means is a popular traditional algorithm that is widely used for text clustering (Karandikar 2010). The algorithm works with distance calculation and assigns each data point to the nearest cluster considering the distance to cluster mean. K-means is used in many text mining tasks to identify the topic distribution in a corpus (Chen et al. 2010). However, k-means as a distance-based method faces the distance concentration problem. Also, it forms spherical shape clusters (Fahim et al. 2008). Density-based approaches are also used to identify topics over documents (Aliguliyev 2009). DBSCAN is a well-known clustering algorithm based on data density (Khan et al. 2014). It is widely used in data mining applications as it is able to handle noise and outliers and, does not require the specified number of clusters initially (Li and Huang 2010; Indah et al. 2019). It is ideal to discover clusters in noisy data sets, and it discovers clusters of arbitrary shape, unlike K-means which generally detects clusters of the round shape. However, for text data that form sparse representation matrices, density-based methods are unable to differentiate clusters based on the density differences.

2.2 Probabilistic topic modelling

Probabilistic latent semantic analysis (PLSA) (Hofmann 1999) was proposed as the first probabilistic topic model. They used a form of dimension reduction that uses a probabilistic model to find the co-occurrence patterns of terms that correspond to semantic topics in a collection of documents.

Latent Dirichlet allocation (LDA) is the first proposed complete generative probabilistic topic modelling technique. LDA models the contributions of different topics to a document by treating each topic as a probability distribution over words and thereby viewing a document as a probabilistic mixture of the associated topics (Blei et al. 2003). The topic modelling research continued to boost with LDA-based topic models that exhibit outstanding performance in many studies with different communicative contexts (Liu et al. 2016). However, all these probabilistic methods which use term counts for the probability calculations are stagnant in the short text that is with the limited number of terms in a document.

2.3 Neural topic modelling

Recently, topic modelling methods that utilize neural variational inference emerged as a new paradigm. Neural topic models use deep neural network architecture to approximate the intractable marginal distribution and thus gain strong generalization ability compared to traditional Bayesian topic models (Wang and Yang 2020). In comparison with traditional topic models, the methods in this paradigm such as the variational autoencoder neural topic models do not require mathematically deriving a new inference algorithm once a change has been done to the model (Srivastava and Sutton 2017).

The initial method proposed under autoencoding variational Bayes-based inference is for LDA and is known as autoencoded variational inference for topic model (AVITM) (Srivastava and Sutton 2017). It has shown promising accuracy results aligning with LDA with much better inference time. The neural autoregressive topic model (DocNADE) is another important method under this category that has shown promising results recently (Larochelle and Lauly 2012). In estimating the probability of observing a new word in a given document given the previously observed words, DocNADE replaces the expensive softmax distribution over words with a hierarchical distribution over paths in a binary tree of words. Though these neural topic models show an extremely efficient inference process, the probability calculation step faces difficulties due to limited words in a short text.

2.4 Matrix factorization-based topic modelling

The origin of topic modelling runs back to latent semantic indexing (LSI)-based models (Papadimitriou et al. 2000). The basis for LSI-based topic model implementation converts high-dimensional, sparse term space to latent semantic space which reflects the topic space through singular value decomposition (SVD)-based matrix factorization (Kherwa and Bansal 2020).

The matrix factorization is a widely used technique in unsupervised learning for dimensionality reduction (Liu et al. 2008). In there, NMF concept (Lee and Sebastian 1999) that matches with natural non-negativity of text data and guarantees the convergence property shows useful wide range of practical applications such as text classification (Buciu 2008), semantic analysis (Lee and Sebastian 1999) and biological data analysis (Pascual-Montano et al. 2006). NMF finds a low-rank approximation of a high-dimensional term–document matrix to identify the topic assignments for documents.

Several variations of NMF were proposed with improvements for effective topic modelling. In Liu et al. (2018), NMF-based model is proposed to learn topics for short texts by directly factorizing a symmetric term correlation matrix. They tried to avoid the sparseness problem in the term–document matrix by constructing a stable and dense matrix representation. The results show the effectiveness of this model in terms of topic visualization, document clustering and document classification. However, since they formulate a quadratic non-convex loss function, the proposed algorithm is not reliable and stable.

The semantics-assisted non-negative matrix factorization (SeaNMF) uses word–context correlation information in the matrix factorization process for short-text topic modelling (Shi et al. 2018). It focuses on giving semantic assistance by factorizing term correlations matrix to enable skip-gram with negative sampling. Recently proposed topic 'regularized non-negative matrix factorization (TRNMF)' (Yi et al. 2020) is another extension of NMF, which uses a regularized non-negative matrix factorization topic model for short texts. To conquer the sparseness problem, they used a pre-trained distributional vector representation. They integrated both word co-occurrence regularization and sentence similarity regularization into the proposed model. This information assists the general NMF process. TRNMF exhibits better results in terms of topic coherence and text classification. However, the computational complexity of the pre-training phase (i.e. modelling word embeddings semantic matrix) is high in TRNMF.

Overall, matrix factorization methods work better for short texts when compared to the probabilistic models. The major drawback of matrix factorization methods is that the factorization only provides an approximation. Therefore, it is

not 100% accurate and results in an information loss within dimensionality reduction. When applying NMF for a corpus of short texts, the term–document matrix becomes extremely sparse, the NMF models may not be accurate in approximating lower-dimensional factor matrices (Virtanen et al. 2008).

While there exists a handful of assistance-driven NMF algorithms for discovering topics in the short text, the application of neighbourhood information on the NMF framework is scarce. Most of the existing methods focus on enabling semantic assistance. To our best knowledge, this is the first method that represents neighbourhood information via Jaccard similarity to couple with NMF process to assist effective topic modelling for short text to minimize the information loss.

3 Methodology

This paper proposes a novel algorithm that couples neighbourhood information with document representation named neighbourhood assistance-based non-negative matrix factorization (NaNMF). NaNMF uses Jaccard similarity to identify neighbourhood affinity for matrix factorization and learns the optimum topic cluster representation in lower-order by iteratively approximating the symmetric document neighbourhood matrix and the document–term representation matrix with minimizing learning error as in Fig. 1.

3.1 Preliminaries

Consider a short document collection $D = \{d_1, d_2, \dots d_N\}$ with N documents and M unique terms. Let each short-text document d_i represents with respective terms as $d_i = \{t_1, t_2, \dots t_l\}$. The text data N in the collection D with the vocabulary of M terms are represented as a document × term matrix in a vector space. Given a non-negative term× document matrix $A \in R^{M \times N}$, and number of topics as lower rank k, standard NMF-based topic modelling methods iteratively find a low-rank approximation to A as a multiplication of two non-negative factor matrices $W \in R^{M \times k}$ and $H \in R^{k \times N}$ as in Eq. 1:

$$A \approx W \times H \tag{1}$$

The term–document matrix A represents each document $d_i \in D$ as a bag of words using term counts as the weights. In contrast to long text, for short text with a limited number of words, term frequency is effective in capturing the inherent concept behind a document.

3.1.1 Neighbourhood-based document affinity

The symmetric matrix $S \in \mathbb{R}^{N \times N}$ is modeled to capture the neighbourhood information via document similarity using the Jaccard coefficient in NaNMF. In S each document pair $d_i, d_j \in D$ is represented with term sets associated with them as in Eq. 2.

$$S_{(d_i,\, d_j)} = \frac{|d_i \cap d_j|}{|d_i \cup d_j|} \tag{2}$$

Jaccard similarity ranges from 0 to 1, resulting in 0 when no shared terms and 1 when both texts are identical. Thus, entries of S range from 0 to 1 where the higher value represents the more similar documents that lay close to each other in higher-dimensional vector space denoting close neighbours. As this measure outputs non-negative values, the document–document affinity matrix constructed in NaNMF contains only non-negative elements, aligning with the non-negativity constraints in NMF.

Then, symmetric NMF is applied to the matrix S to generate two commutative matrices H and $H_n \in R^{k \times N}$ as in Eq. 3.

$$S \approx H^T H_n \tag{3}$$

3.1.2 NMF framework

We conjecture that the use of both the commonalities among the terms based on the document in which they appear and the commonalities among the documents based on the terms which they share can assist short-text topic clustering to achieve better performance. Thus, NaNMF combines the latent neighbourhood information within the document affinity matrix S with the matrix A which includes inter-document relationships with associated terms within the objective function. It uses both of these pieces of information to learn a common document×topic matrix H as in Eq. 4 minimizing learning error.

$$\min_{W,H \geq 0} \|A - WH\|_F + \min_{H,H_n \geq 0} \|S - H^T H_n\|_F \tag{4}$$

Thereby, NaNMF is able to effectively incorporate the neighbourhood-based assistance using matrix S to the NMF model for compensating the information loss in dimensionality reduction with matrix A.

3.1.3 Solving optimization problem

We have employed the block coordinate descent (BCD) optimization concept (Tseng 2001) to optimize the objective function in Eq. 4 which is a widely used algorithm for solving both convex and non-convex problems in the form of multiple blocks of variables. It divides the matrix members into several disjoint subgroups and iteratively minimizes the objective function with respect to the members of each topic $c_j \in C$ at a time where $C = \{c_1, c_2 \dots c_k\}$.

BCD computes sub-problems that depend on each other sequentially and uses the most recent values of the associated factor matrices for updating the following factor matrices. NaNMF set members to be random non-negative values at the initialization. Firstly, the BCD update rule finds W using the term–document matrix A and the initial matrix H_n as in Eq. 5.

$$W_{(:,j)} \Leftarrow \left[W_{(:,j)} + \frac{\left(AH_n\right)_{(:,j)} - \left(WH_n^T H_n\right)_{(:,j)}}{\left(H_n^T H_n\right)_{(j,j)}} \right]_+ \tag{5}$$

Secondly, the matrix H_n is updated using the current values of W and other members as in Eq. 6:

$$H_{n(:,j)} \leftarrow \left[H_{n(:,j)} + \frac{\left(A^T W\right)_{(:,j)} + \left(SH_n\right)_{(:,j)}}{\left(W^T W\right)_{(j,j)} + \left(H_n^T H_n\right)_{(j,j)}} - \frac{\left(H_n W^T W\right)_{(:,j)} + \left(H_n H_n^T H_n\right)_{(:,j)}}{\left(W^T W\right)_{(j,j)} + \left(H_n^T H_n\right)_{(j,j)}} \right]_+ \tag{6}$$

Then, S, the matrix representing neighbourhood-based affinity and most recent values of H_n are used for updating H as in Eq. 7:

$$H_{(:,j)} \Leftarrow \left[H_{(:,j)} + \frac{\left(S^T H_n\right)_{(:,j)} - \left(HH_n^T H_n\right)_{(:,j)}}{\left(H_n^T H_n\right)_{(j,j)}} \right]_+ \tag{7}$$

Iteratively, W, H_n and H are updated until convergence using the above update rules aligning with BCD update rules.

3.2 Topic assignment

Usually, short text like tweets have only a few words and semantically belong to a single topic in comparison with long-text documents which are a mixture of topics (Zhao et al. 2011). Thus, the final topic assignment for each document is obtained by applying hard clustering on H. H represents the likelihood coefficients of each document being assigned to each topic $c_j \in C$.

$$H_F = \text{argmax} \sum_{j=1}^{k} \left(H_{(:,j)}\right) \tag{8}$$

NaNMF chooses a topic that possesses the highest coefficient within H as the topic assigned to a specific document and forms vector H_F as in Eq. 8. The complete NaNMF algorithm is given in Algorithm 1.

Table 1 Data set properties

Data set	Size	Vocabulary size	Average length	Clusters	Density
Cancer	13,002	4552	15.598	5	0.002105
Health	12,101	4683	15.731	4	0.002101
Sports	13,946	5091	14.992	8	0.001859

Algorithm 1 Algorithm for the proposed NaNMF

Require: $W \geq 0, H \geq 0, H_n \geq 0$ with random real numbers and number of topics k

```
while convergence of Eq.4 do
    for j = 1 to k do
        Compute W_(:,j) by Eq.5
        Compute H_n(:,j) by Eq.6
        Compute H_(:,j) by Eq.7
    end for
end while
H_F = argmax Σ_{j=1}^{k} (H_(:,j))
```

4 Empirical analysis

In this section, we present the experimental evaluation of the proposed NaNMF algorithm. Experiments were done using Python 3.5 on a single processor of 1.2 GHz Intel (i7) with an 8 GB memory. First, we present the description of the real-world data sets used, pre-processing steps and the standard evaluation measures used to determine the accuracy of topic detection methods.

4.1 Data sets

We used three Twitter data sets obtained from (Mohotti and Nayak 2018) spanning across cancer, health and sports domains, which consist of different numbers of topics as reported in Table 1 for quantitative evaluation. Usually, a tweet document has only 280 characters, and thereby considered data sets are extremely sparse as given in the density column. These data sets contained the ground truth labels to benchmark the algorithmic outcome.

Additionally, two other Twitter data sets spanning 'Organic Food' and 'COVID-19' listed below have been used in the case studies for qualitative evaluation.

1. *COVID-19* The COVID-19 tweets data set is collected using the 'Covid' keyword, posted on 2021-07-06 using public Twitter API (Size: 9470 tweets)
2. *Organic Food* This data set is collected using the keyword 'Organic Food' within the period of 2020-01-01 until 2020-11-01 (Size: 8842 tweets) (Mohotti et al. 2021).

4.2 Pre-processing

Data sets have been prepossessed for word lowercasing, punctuation removal, stop word removal and tokenization following the standard text pre-processing approach. Additionally, since we are working with noisy short-text data, we applied special pre-processing techniques which are emoji removal, username replacement, hash symbol replacement and number removal.

Then, the matrix A is presented as a vector space model (VSM) with the term frequency as weighting and the symmetric matrix S is presented with Jaccard similarity values between document pairs.

4.3 Baselines

As primary baselines, we have chosen unsupervised topic detection methods from the major categories in the existing literature to compare against the proposed unsupervised NaNMF as listed below:

1. *Latent Dirichlet allocation (LDA)* (Blei et al. 2003) LDA is a probabilistic approach to retrieve topics given the number of topics k.
2. *Non-negative matrix factorization (NMF)* (Lee and Sebastian 1999) Matrix decomposition-based dimension reduction and clustering approach to retrieve topics simultaneously given the number of topics k.
3. *Semantics-assisted non-negative matrix factorization (SeaNMF)* (Shi et al. 2018) SeaNMF is a short-text topic modelling method with semantic assistance and NMF was given the number of topics k.
4. *Biterm topic model (BTM)* (Yan et al. 2013) BTM is short-text generative topic modelling method, and it learns the topics by generating word co-occurrence patterns in the whole corpus given the number of topics k.
5. *Autoencoding variational inference for topic model (AVITM)*[1] (Srivastava and Sutton 2017) AVITM-ProdLDA is a neural topic model that uses autoencoded variational Bayes inference algorithm for Latent Dirichlet allocation .

[1] https://github.com/akashgit/autoencoding_vi_for_topic_models.

In addition, ablation studies have been performed with variants of NaNMF using various neighbourhood assistance and document representation techniques.

The experiments were conducted to find the most effective neighbourhood modelling technique considering (1) **Cosine similarity** (cosine) which measures correlation as the cosine angle between the term vectors (Huang 2008), (2) **Sigmoid Jaccard similarity** (Sigmoid) which is an improved version of the Jaccard similarity based on features (Likavec et al. 2019), (3) **Word2Vec similarity** (Word2Vec-Sim) that capture document similarity via semantic distance between words (Handler 2014).

The next set of experiments was conducted to find out the best term–document representation using different term weighting techniques: (1) **TF-IDF** that considers how relevant a word is to a document based on frequency in a collection of documents (Liu et al. 2018) and (2) **IDF** that considers how rare a word is across documents (Liu et al. 2018).

4.4 Evaluation measures

Quantitative evaluation measures which evaluate based on numeric values are categorized as external and internal measures (Rendon et al. 2011). External measures use ground truth labels to check the accuracy, while internal measures use the information obtained within the topic clustering process itself such as cluster tightness (Rendon et al. 2011).

The accuracy of the topic clusters in NaNMF is evaluated by the standard pairwise F1-score which calculates the harmonic average of precision and recall, and normalized mutual information (NMI) which measures the purity against the number of clusters using ground truth labels. Additionally, topic coherence which employs the pointwise mutual information (PMI) to measure the coherence of topics is used. Additionally, we used word clouds (Heimerl et al. 2014) to qualitatively evaluate the accuracy of the NaNMF.

4.5 Quantitative analysis

We evaluate NaNMF to show its effectiveness quantitatively in this section against relevant baselines using ground truths and internal measures. Also, this section covers the sensitivity of the proposed NaNMF.

4.5.1 Accuracy analysis

Results with extrinsic measures against other topic modelling Table 2 shows accuracy comparison against baselines using ground truths. These results show how novel coupled matrix factorization with neighbourhood-based assistance in NaNMF accurately identifies topic structures in the short text superior to the other baselines.

Table 2 Results against existing baselines

Baseline	Metric	Cancer	Health	Sports	Average
LDA	F1	0.26	0.28	0.23	0.26
	NMI	0.07	0.04	0.16	0.09
	Topic coherence	0.56	0.55	0.61	0.57
NMF	F1	0.79	0.84	0.96	0.86
	NMI	0.74	0.80	0.95	0.83
	Topic coherence	0.64	0.63	0.69	0.65
Biterm	F1	0.66	0.84	0.75	0.75
	NMI	0.57	0.75	0.74	0.69
	Topic coherence	0.45	0.47	0.53	0.48
SeaNMF	F1	0.97	0.68	0.88	0.84
	NMI	0.95	0.60	0.88	0.81
	Topic coherence	0.75	0.83	0.70	0.76
AVITM	F1	0.20	0.26	0.13	0.20
	NMI	0.00	0.00	0.13	0.04
	Topic coherence	0.92	0.93	0.94	0.93
NaNMF	F1	**0.98**	**0.99**	**0.97**	**0.98**
	NMI	**0.96**	**0.97**	**0.96**	**0.96**
	Topic coherence	**0.81**	**0.80**	**0.72**	**0.78**

The performance of NaNMF (Our proposed method) are given in bold

As reported in Table 2, NaNMF gives the best results on average with F1-score and NMI, confirming the effectiveness of coupling neighbourhood-based assistance with document–term representation in the NMF process. Applying the traditional LDA model for short texts lacks sufficient information for effective statistical learning due to the extremely sparse corpus. Thus, the results are significantly inferior when compared with other baselines. The Biterm topic model, which is also an extended LDA-based probabilistic model, exhibits better results than LDA as it works by directly modelling the generation of word co-occurrence patterns. However, all the matrix factorization-based models outperformed BTM due to their characteristic short length which challenges probability calculations.

On average, NMF and SeaNMF results are close; however, data set-wise they differ. SeaNMF results in higher accuracy for the Cancer data set given its property of lower vocabulary size. SeaNMF is an improved version of NMF with provided semantic assistance, while NaNMF provides neighbourhood-based assistance to NMF. When comparing SeaNMF and NaNMF, we can observe that NMF which utilizes neighbourhood-based assistance simply with Jaccard performs better than employing semantic assistance through Skip Gram with Negative Sampling concept considering term co-occurrences.

Neural topic models such as AVITM are based on neural network concepts in learning topic distributions. Specifically, AVITM (AVITM-ProdLDA) which generally shows a similar performance more efficiently compared to LDA shows an inferior performance in our data sets in terms of F1-score and NMI when checked with the labels statistically. The limited number of words in the short text leads to this poor clustering quality which propagates forward in the neural learning process.

In summary, NaNMF that captures neighbourhood information via Jaccard similarity assists the information loss resulting in higher-to-lower-dimensional projection effectively. Thereby, NaNMF attains higher accuracy in topic modelling as confirmed by extrinsic measures.

Results with intrinsic measures against other topic modelling methods

The closeness of topic words measured in terms of topic coherence also confirms that NaNMF outperforms most of the other baselines as per Table 2. It shows the superiority of NMF-based models for short text in comparison with traditional probabilistic models in short text with limited terms. Additionally, these results validate the success of assistance-based NMF models with higher results for SeaNMF and NaNMF.

However, it is interesting to note that neural topic models show better interpretability in resulted topics. The topic coherence that calculates the degree of semantic similarity between words in the topics confirms this superiority of AVITM as per the results in Table 2. However, we cannot guarantee the prediction power of intrinsic evaluation as a whole over extrinsic evaluation (Chiu et al. 2016).

In general, both internal and external measures show the accuracy of the proposed neighbourhood-based assistance concept in NaNMF.

Results against other neighbourhood affinity modelling

Table 3 shows the accuracy results gained from different neighbourhood-based assistance approaches via

Table 3 Results against different similarity measures

	Jaccard(NaNMF)		Cosine		Sigmoid		Word2Vec	
	F1	NMI	F1	NMI	F1	NMI	F1	NMI
Cancer	**0.98**	**0.96**	0.97	0.95	0.98	0.95	0.98	0.96
Health	**0.99**	**0.97**	0.97	0.94	0.98	0.95	0.99	0.97
Sports	**0.97**	**0.96**	0.89	0.90	0.97	0.96	0.97	0.96
Average	**0.98**	**0.96**	0.95	0.93	0.97	0.95	0.98	0.96

The performance of NaNMF (Our proposed method) are given in bold

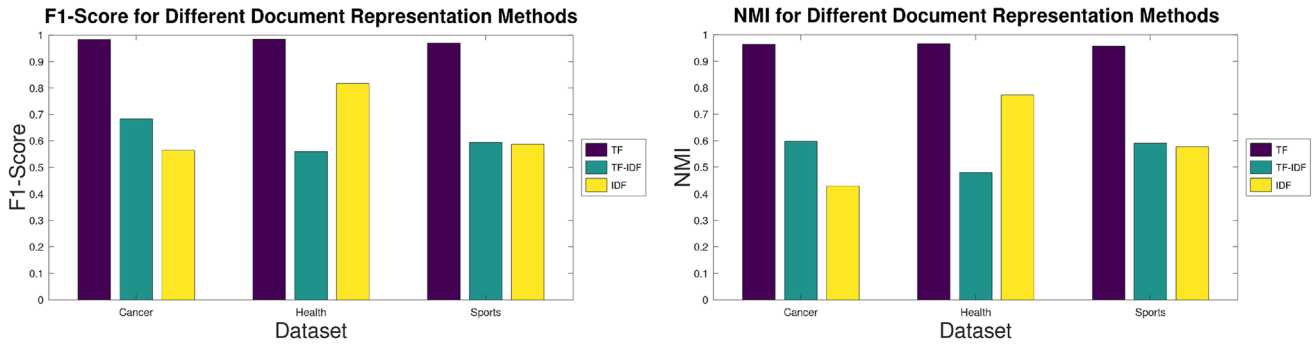


Fig. 2 F1-Score and NMI for different document representation methods

text similarity measures employed to compensate for the information loss. It highlights that NaNMF which uses the standard Jaccard similarity is more accurate in modelling the neighbourhood information in comparison with other baseline methods except for Word2VecSim.

Cosine similarity scores the lowest among all the considered similarity measures. It computes the cosine angle between two document representation vectors. However, the low word co-occurrence in short texts causes this measure to be inferior. Although Sigmoid Jaccard is an improved version of Jaccard similarity, it falls short compared to the Jaccard similarity in NaNMF. Its feature-based concept is not suited to short texts that lack features due to the limited length.

Word2Vec for neighbourhood assistance exhibits almost similar results to Jaccard similarity in NaNMF. It uses a pre-trained corpus to determine text similarities though it captures context considering semantic similarity as well. This reveals an interesting observation; neighbourhood information in higher-dimension could be captured considering semantic information. However, the computational complexity of Word2VecSim is significantly high in contrast to Jaccard similarity in NaNMF, especially for large data sets due to the requirement of pre-training of the corpus to identify similarities. Jaccard similarity exhibits excellent results simply as it iterates over each word in both texts when calculating pairwise similarity. Thus, simple and efficient Jaccard similarity in NaNMF shows its effectiveness as a better technique to model neighbourhood information in assisting the NMF process.

Results against other term weighting methods

Figure 2 shows the effectiveness of the type of text representation scheme used in NaNMF against other representation methods, TF-IDF and IDF representation.

In NaNMF, a document is represented as a bag of words where each word is given the weight using its count (term frequency (TF)) in the document. In TF-IDF (term frequency-inverse document frequency), weight values determined the relevancy of a word to a document in a collection, while in IDF (inverse document frequency), values measure how rare the terms are in documents. Results show that utilizing TF for short document representation achieves superior results compared to other TF-IDF and IDF representation models in terms of both F1-score and NMI results. Additionally, TF representation cooperates with the coupled matrix factorization process in NaNMF.

4.5.2 Sensitivity analysis

The proposed NaNMF is for short-text topic modelling and experimental analysis is performed on Twitter. Tweets contained a lot of noise and unstructured text phrases. Thus, it requires additional pre-processing that goes beyond standard text pre-processing techniques.

Table 4 reports the results of the proposed algorithm with standard text pre-processing techniques and after applying special text cleaning techniques as in NaNMF. Tweets are naturally noisy as they contain lots of unstructured phrases and symbols in the forms of emojis, hashtags, usernames and special characters which do not play an important role in topic modelling. Results show that special pre-processing techniques eliminate the unnecessary noise from the Twitter data sets and thereby NaNMF is able to achieve significantly accurate results.

To achieve the optimized solution for NaNMF, the learning error of the objective function is iteratively minimized during factorization over 100 cycles. Convergence

Table 4 Experimental results with special pre-processing and standard text pre-processing

	Standard		Special (NaNMF)	
	F1	NMI	F1	NMI
Cancer	0.69	0.63	0.98	0.96
Health	0.83	0.77	0.99	0.97
Sports	0.89	0.88	0.97	0.96
Average	0.80	0.76	0.98	0.96

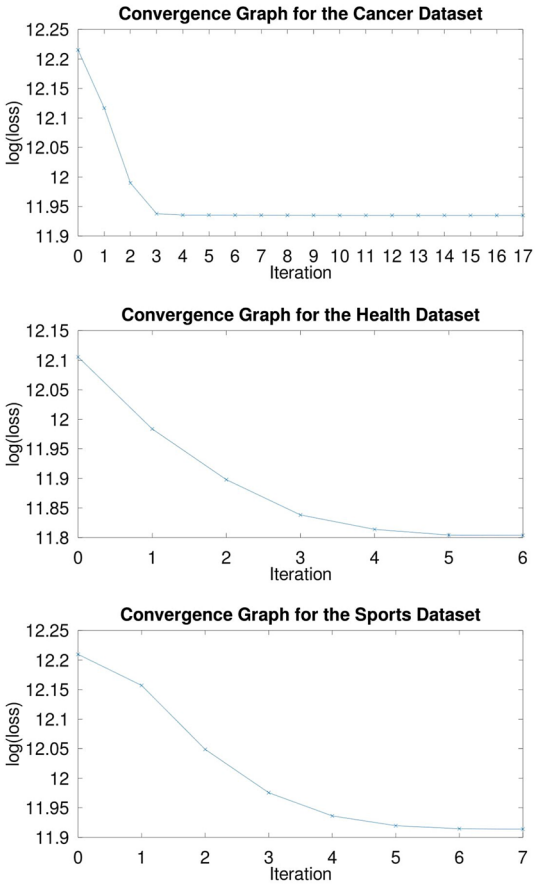


Fig. 3 Convergence of NaNMF Algorithm

of some NMF algorithms is not guaranteed as they depend on the update rules (Takahashi and Hibi 2014). Thus, the convergence of objective function/optimization validates the correctness of update rules and thereby the reliability. We experimentally tested the convergence of the proposed algorithm NaNMF. The graphs shown in Fig. 3, guarantee that the NaNMF algorithm reaches at least a local minimum through the defined update rules corresponding to the optimization algorithm.

4.5.3 Complexity and scalability analysis

The incremental sampling on the Sports data set collection is used to demonstrate the scalability of NaNMF. Generally, NaNMF has $O(n^2)$ computational complexity where n is the size of the data set as it decomposes document representation matrices. This case is similar to all the NMF-based methods (Shi et al. 2018; Gillis 2011). However, this polynomial-time complexity results in a higher time requirement for larger data sets for NaNMF though it does not show exponential-time growth. LDA or other neural topic modelling methods also show similar or higher computational

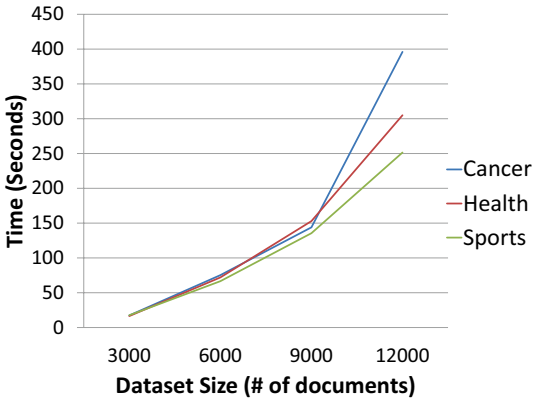


Fig. 4 Scalability of NaNMF Algorithm

complexity, so they do require higher time for larger data sets (Huang et al. 2002; Livni et al. 2014).

Figure 4 shows the trend NaNMF exhibits with the time taken against the size of the corpus. It confirms that with the incremental samples, the time taken is close to a near-linear pattern though NaNMF shows a quadratic complexity.

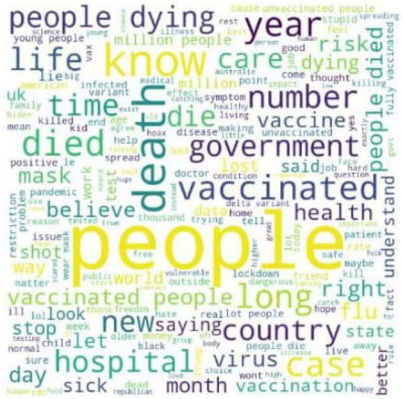
4.6 Qualitative analysis

Qualitative analysis is performed with case studies to validate the NaNMF in real-world settings. The word cloud visualization is employed to have an insight into the obtained topics. The manual interpretation of the word clouds in two case studies shows the meaningfulness of the derived topics in the COVID-19 domain and Organic Food domain. As we used real-world, unlabeled data sets, we had to determine the number of topics k in advance. The average topic coherence informs the tightness of topic clusters with a numerical value for a given number of topics k. Thus, the number of topics with the highest average topic coherence is chosen as the exact number of topics (i.e. k) in these case studies aligning with the standard practice (Chiang and Mirkin 2010) for both case studies. In both case studies, k resulted as 8 by the coherence.

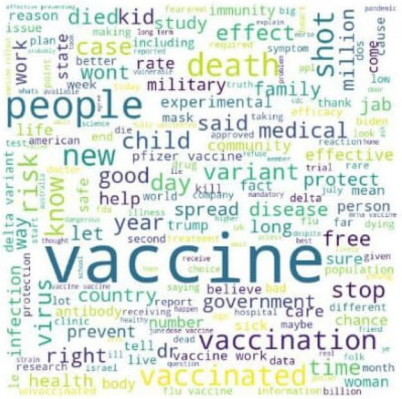
4.6.1 Case study I: COVID-19

The topics revealed by this COVID-19 case study are as follows: (1) people-related concerns, (2) vaccine, (3) new cases, (4) work–life, (5) death causes/rate, (6) delta and other variants, (7) risks and symptoms, and (8) pandemic time, as shown in Fig. 5.

The topic **People-related concerns** in Fig. 5a includes a general discussion that shows the impact of corona on the people such as the vaccinated people, hospitals for people, people dying due to coronavirus, and death numbers together with the involvement of governments to control

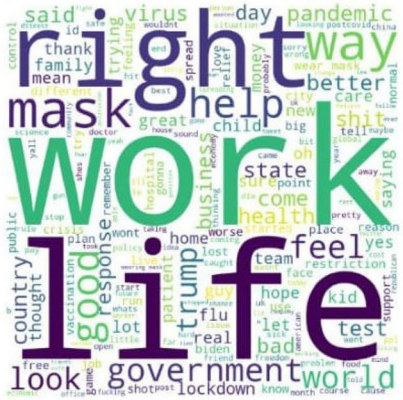


A: Case Study 1 – Word Cloud 1 (People related concerns)

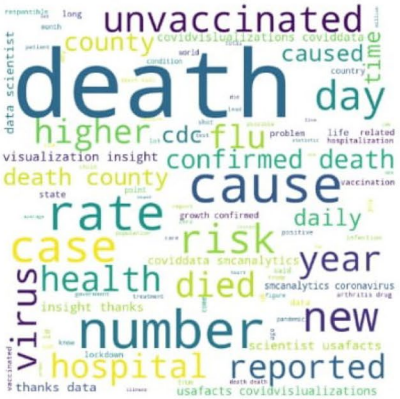


B: Case Study 1 – Word Cloud 2 (Vaccines)

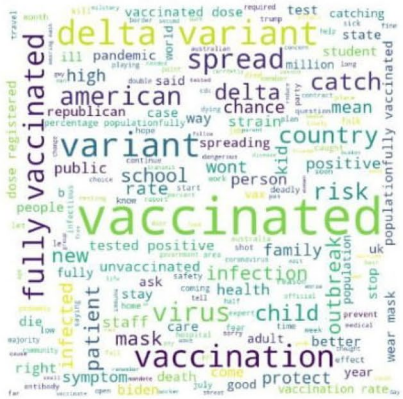
C: Case Study 1 – Word Cloud 3 (New Cases)



D: Case Study 1 – Word Cloud 4 (Work-life)



E: Case Study 1 – Word Cloud 5 (Death cause/rate)



F: Case Study 1 – Word Cloud 6 (Delta and other variants)

G: Case Study 1 – Word Cloud 7 (Long-term risks and symptoms)

H: Case Study 1 – Word Cloud 8 (Pandemic time)

Fig. 5 Word clouds of case study I

this situation. The next topic **Vaccine** in Fig. 5b covers specific vaccine-related concerns such as vaccination process, government initiatives, immunity, number of shots, vaccine efficacy against variants, and the risks. The topic **New case** in Fig. 5c shows topic words on the new cases and associated statistics with the discussions about new cases on a daily and weekly basis and months that show outbreaks.

Figure 5d shows the topic of **Work–life** that talks about work–life and work rights during this COVID-19 and how it has influenced work–life with wearing masks and government lockdown. The next topic **Death causes/rate** in Fig. 5e focuses on COVID-19 deaths. The included topic words cover discussion on the reasons behind the confirmed number of deaths and death rate among the unvaccinated communities. Then, the topic of **Delta and other variants** in Fig. 5f focuses on the vaccination against the spread of the new delta variant. The topic words show discussions about the effectiveness of the vaccination against the variants and differences among fully vaccinated, vaccinated, and unvaccinated cases.

The topic **Risks and symptoms** in Fig. 5g concentrates on symptoms, risks and the long-term effects of the virus. The last topic **Pandemic time** in Fig. 5h shows how the COVID-19 pandemic has changed day-to-day life, work and school education together with the people's experience at home with kids in the lockdown.

The meaningfulness of these revealed topics in the COVID-19 domain through the included topic words confirms the validity of NaNMF for short-text topic modelling on the Twitter platform.

4.6.2 Case study II: organic food

This case study revealed meaningful topics and topic words that aligned with theories in buying behaviour of organic food (Mohotti et al. 2021). The identified topics by NaNMF are (1) healthy food, (2) health benefits, (3) food services, (4) vegan/gluten-free food, (5) nutritional facts, (6) fresh products, (7) organic gardening and (8) organic farming in the organic food domain. They are able to confirm the reliability of the NaNMF with their relevance to the domain.

The topic **Healthy Food** in Fig. 6a talks about consuming natural and organic food for a healthy lifestyle. It also talks about the deliciousness of natural organic products. Then, **Healthy benefits** in Fig. 6b covers topic words covering the benefits of eating an organic diet with fewer pesticides and high in nutrition. The next topic **Food services** in Fig. 6c covers talks about food services related to organic food. Then, **Vegan/Gluten-free Food** in Fig. 6d covers another belief goes with organic food; organic food for a vegan, gluten-free or healthy lifestyle.

Figure 6e with topic **Nutritional facts** covers topic words related to the nutritional facts of organic food such as their caloric values, fibre content, protein content, and fat content. The next topic **Fresh products** in Fig. 6f talks about the freshness factor in organic buying. The last two topics of **Organic gardening** in Fig. 6g and **Organic farming** in Fig. 6h discuss which plants are suitable for organic gardening and sustainable organic farming with natural fertilizer and fewer pesticides.

5 Conclusion

This paper proposes an accurate topic modelling algorithm for short texts, to deal with higher dimensionality and associated extreme sparseness in the short-text representation. We identified that traditional topic modelling methods face challenges in handling higher feature dimensions in text data. Additionally, probabilistic topic modelling methods fail in probability estimation for short texts due to their lack of word co-occurrence. This issue leads to an information loss in dimensionality reduction-based methods such as NMF. As a solution, we introduced a novel topic modelling algorithm called NaNMF, which was inspired by the conjecture that assistance-based NMF methods can outperform existing standard methods for short-text topic modelling. We leveraged neighbourhood-based information to compensate for the information loss happening in NMF during the higher-to-lower-dimensional approximation process.

In summary, the proposed model successfully uses document–document similarity information which is calculated using the Jaccard similarity coefficient to provide neighbourhood-based assistance to the model. An extensive experimental study has been conducted with both qualitative and quantitative evaluation. The performance against existing state-of-the-art methods on three real twitter data sets illustrates that NaNMF outperforms other methods with higher accuracy and topic coherence. Also, case study analysis using data sets crawled from Twitter API shows its potential in identifying meaningful topic clusters for real-world short-text data.

This article presents substantial work in the area of short-text topic modelling. However, identifying topic distribution over a tweet is important for trend detection. Therefore, future directions focus on inter-topic modelling and finding topics for an individual tweet. Also in the future, we will explore how to efficiently combine semantic information with the neighbourhood concept in short-text topic modelling.

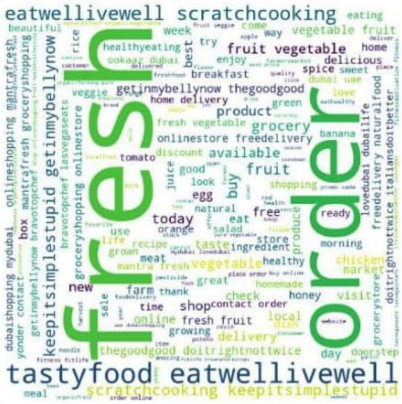
A: Case Study 2 – Word Cloud 1 (Healthy food)

B: Case Study 2 – Word Cloud 2 (Health benefits)

C: Case Study 2 – Word Cloud 3 (Food services)

D: Case Study 2 – Word Cloud 4 (Vegan, gluten-free food)

E: Case Study 2 – Word Cloud 5 (Nutritional facts)



F: Case Study 2 – Word Cloud 6 (Fresh products)

G: Case Study 2 - Word Cloud 7 (Organic gardening)

H: Case Study 2 - Word Cloud 8 (Organic farming)

Fig. 6 Word clouds of case study II

References

Aggarwal CC, Zhai C (2012) An introduction to text mining. In: Mining text data. Springer, pp1–10

Aliguliyev RM (2009) Information sciences. Perf Eval Dens-Based Clust Methods 179(20):3583–3602

Balasubramaniam T, Nayak R, Luong K, Bashar MA (2021) Identifying covid-19 misinformation tweets and learning their spatio-temporal topic dynamics using nonnegative coupled matrix tensor factorization. Soc Netw Anal Min 11(1):1–19

Balbi S (2010) Beyond the curse of multidimensionality: high dimensional clustering in text mining. Statistica Applicata-Ital J Appl Stat 22(1):53–63

Blei DM (2012) Probabilistic topic models. Commun ACM 55(4):77–84

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Buciu I (2008) Non-negative matrix factorization, a new tool for feature extraction: theory and applications. Int J Comput Commun Control 3(3):67–74

Chen Y, Qin B, Liu T, Liu Y, Li S (2010) The comparison of som and k-means for text clustering. Comput Inf Sci 3(2):268–274

Chiang MM-T, Mirkin B (2010) Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. J Classif 27(1):3–40

Chiu B, Korhonen A, Pyysalo S (2016) Intrinsic evaluation of word vectors fails to predict extrinsic performance. In: Proceedings of the 1st workshop on evaluating vector-space representations for NLP, pp 1–6

Fahim AM, Saake G, Salem AM, Torkey FA, Ramadan MA (2008) K-means for spherical clusters with large variance in sizes. J World Acad Sci Eng Technol 35:177–182

Ferdous R et al. (2009) An efficient k-means algorithm integrated with jaccard distance measure for document clustering. In: 2009 first asian himalayas international conference on internet. IEEE, pp 1–6

Gillis N et al. (2011) Nonnegative matrix factorization: Complexity, algorithms and applications. Unpublished doctoral dissertation, Université catholique de Louvain. Louvain-La-Neuve: CORE

Gomaa WH, Fahmy AA et al (2013) A survey of text similarity approaches. Int J Comput Appl 68(13):13–18

Handler A (2014) An empirical study of semantic similarity in wordnet and word2vec. Master's thesis, University of New Orleans, USA

Heimerl F, Lohmann S, Lange S, Ertl T (2014) Word cloud explorer: Text analytics based on word clouds. In: *2014 47th Hawaii international conference on system sciences*. IEEE, pp 1833–1842

Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp 50–57

Huang A et al. (2008) Similarity measures for text document clustering. In: Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, vol 4, pp 9–56

Huang J, Peng M, Li P, Zhiwei H, Chao X (2020) Improving biterm topic model with word embeddings. World Wide Web 23(6):3099–3124

Huang R, Liu Q, Lu H, Ma S (2002) Solving the small sample size problem of lda. In: *2002 international conference on pattern recognition*. IEEE, vol 3, pp 29–32

Indah RNG, Novita R, Kharisma OB, Vebrianto R, Sanjaya S, Andriani T, Sari WP, Novita Y, Rahim R et al. (2019) Dbscan algorithm: twitter text clustering of trend topic pilkada pekanbaru. In: Journal of Physics: Conference Series. IOP Publishing, vol 1363, pp 012001

Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L (2019) Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. Multimed Tools Appl 78(11):15169–15211

Karandikar A (2010) Clustering short status messages: a topic model based approach. Master's thesis, University of Maryland, USA

Khan K, Rehman SU, Aziz K, Fong S, Sarasvady S (2014) Dbscan: past, present and future. In: The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014). IEEE, pp 232–238

Kherwa P, Bansal P (2020) Topic modeling: a comprehensive review. EAI Endors Trans Scal Inf Syst, 7(24)

Köppen M (2000) The curse of dimensionality. In: 5th online world conference on soft computing in industrial applications (WSC5), vol 1, pp 4–8

Larochelle H, Lauly S (2012) A neural autoregressive topic model. Adv Neural Inf Process Syst 25

Lee DD, Sebastian SH (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788–791

Li Q, Huang X (2010) Research on text clustering algorithms. In: *2010 2nd international workshop on database technology and applications*. IEEE, pp 1–3

Likavec S, Lombardi I, Cena F (2019) Sigmoid similarity-a new feature-based similarity measure. Inf Sci 481:203–218

Liu C-z, Sheng Y-x, Wei Z-q, Yang Y-Q (2018) Research of text classification based on improved tf-idf algorithm. In: 2018 IEEE international conference of intelligent robotic and control engineering (IRCE). IEEE, pp 218–222

Liu L, Tang L, Dong W, Yao S, Zhou W (2016) An overview of topic modeling and its current applications in bioinformatics. SpringerPlus 5(1):1–22

Liu W, Yuan K, Ye D (2008) Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis. J Biomed Inform 41(4):602–606

Livni R, Shalev-Shwartz S, Shamir O (2014) On the computational efficiency of training neural networks. Adv Neural Inf Process Syst 27

Mohotti WA, Mohotti NK, Wang D, Soontiens W (2021) Driving forces behind organic food through topic modelling on social networks. In: 2021 international conference on multidisciplinary approaches in science (ICMAS), p 85

Mohotti WA(2020) Unsupervised text mining: effective similarity calculation with ranking and matrix factorization. PhD thesis, Queensland University of Technology

Mohotti WA, Nayak R (2018) Corpus-based augmented media posts with density-based clustering for community detection. In: 2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI). IEEE, pp 379–386

Muflikhah L, Baharudin B (2009) Document clustering using concept space and cosine similarity measurement. In: 2009 international conference on computer technology and development. IEEE, vol 1, pp 58–62

Muthu B, Cb S, Kumar PM, Kadry SN, Hsu CH, Sanjuan O, Crespo RG (2021) A framework for extractive text summarization based on deep learning modified neural network classifier. Trans Asian Low-Resour Lang Inf Process 20(3):1–20

Papadimitriou CH, Prabhakar R, Tamaki H, Vempala S (2000) Latent semantic indexing: a probabilistic analysis. J Comput Syst Sci 61(2):217–235

Pascual-Montano A, Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Marqui RD (2006) bionmf: a versatile tool for non-negative matrix factorization in biology. BMC Bioinform 7(1):1–9

Rendon E, Abundez I, Arizmendi A, Quiroz EM (2011) Internal versus external cluster validation indexes. Int J Comput Commun 5(1):27–34

Salloum SA, Al-Emran M, Abdel Monem A, Shaalan K (2017) A survey of text mining in social media: facebook and twitter perspectives. Adv Sci Technol Eng Syst J 2(1):127–133

Sheikh TH (2017) Text mining and its applications. Int J Allied Pract Res Rev 4(11):1–8

Shi T, Kang K, Choo J, Reddy CK (2018) Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In: Proceedings of the 2018 World Wide Web conference, pp 1105–1114

Srivastava A, Sutton C (2017) Autoencoding variational inference for topic models. Preprint arXiv:1703.01488

Takahashi N, Hibi R (2014) Global convergence of modified multiplicative updates for nonnegative matrix factorization. Comput Optim Appl 57(2):417–440

Takeuchi K, Ishiguro K, Kimura A, Sawada H (2013) Non-negative multiple matrix factorization. In: Twenty-third international joint conference on artificial intelligence

Tseng P (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. J Optim Theory Appl 109(3):475–494

Virtanen T, Cemgil AT, Godsill S (2008) Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In: 2008 IEEE international conference on acoustics, speech and signal processing. IEEE, pp 1825–1828

Wang X, Yang Y (2020) Neural topic model with attention for supervised learning. In: International conference on artificial intelligence and statistics. PMLR, pp 1147–1156

Wang Z, Cheng J, Wang H, Wen J (2016) Short text understanding: a survey. J Comput Res Dev 53(2):262

Wu S, Liu F, Zhang K (2020) Short text similarity calculation based on jaccard and semantic mixture. In: International conference on bio-inspired computing: theories and applications. Springer, pp 37–45

Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp 267–273

Yangyang X, Yin W (2017) A globally convergent algorithm for non-convex optimization based on block coordinate update. J Sci Comput 72(2):700–734

Yan X, Guo J, Lan Y, Cheng X (2013) A biterm topic model for short texts. In: Proceedings of the 22nd international conference on World Wide Web, pp 1445–1456

Yan X, Guo J, Liu S, Cheng X-q, Wang Y (2012) Clustering short text using ncut-weighted non-negative matrix factorization. In: Proceedings of the 21st ACM international conference on Information and knowledge management, pp 2259–2262

Yi F, Jiang B, Jianjun W (2020) Topic modeling for short texts via word embedding and document correlation. IEEE Access 8:30692–30705

Zhao WX, Jiang J, Weng J, He J, Lim E-P, Yan H, Li X (2011) Comparing twitter and traditional media using topic models. In: *European conference on information retrieval*. Springer, pp 338–349