Expert Systems With Applications 193 (2022) 116475

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Email thread sentiment sequence identification using PLSA clustering algorithm

Ulligaddala Srinivasarao *, Aakanksha Sharaff

Department of Computer Science and Engineering, National Institute of Technology Raipur, Chhattisgarh, India

ARTICLE INFO

Keywords:
Sentiment email clusters
Probabilistic latent semantic analysis
Topic modeling
SWN lexicon
Sentiment sequence of threads

ABSTRACT

Email messaging is the most common way of providing effective communication between internauts. Consequently, the total sent and received emails count will be increased. But, the internaut can't remember all such emails. Even though email thread identification approaches give satisfactory benefits to the internauts, but they may fail to alert them for a cause to identify the sentiments behind an email thread. To address, this issue Probabilistic Latent Semantic Analysis clustering algorithm has been used in this paper to identify the email sentiment thread sequence. The sentiment and the thread sequence within the emails have been discovered as clustering sentiment polarity and temporal categories with the help of PLSA clusters. At the initial stage, we used three feature extraction methods, latent semantic analysis (LDA), bag of words (BoW), TF-IDF and SentiWordNet (SWN) lexicon for generating sentiment features of email. Next, Probabilistic Latent Semantic Analysis algorithm is used to form email clusters based on sentiment features. Thus, it helps to identify thread sentiment and sequence of sentiment threads. Email threads give a mechanism by which any user will be able to find out the sequence in the thread on the basis of sentiment analysis of email related to a specific set of communication during a specific time period. Various parameters evaluation measures have been considered in this work to evaluate the proposed model such as accuracy, precision, recall and F-measure, and the proposed algorithm is compared with other standard algorithms. Furthermore, a statistical test has also been performed.

1. Introduction

Generally, Sentiment Analysis (SA) is applicable for multiple problems, which consists of several sub-problems such as aspect-based extraction, feature extraction-based sentiment classification tasks. Sentiment analysis has been categorized into three classification problems viz. aspect-level, sentence-level and document-level (Liu, 2015). Among the three classification problems, document sentiment analysis has been considered as the most crucial and fundamental granularity because it will extract the sentiments or opinions from the whole document (Tang, Qin, & Liu, 2015). The sequence within the documents should be identified between the characters for feature orders to identify the sentiments of a document correctly. The essential characteristics of email are being recognized as a topic-oriented communication system, official language as well as its essential nature. Considering email for research is a good choice compared to other social media network data, such as Amazon reviews and micro-blog information because the length of review data and micro-blog is restricted by character limitations whereas, email data is relatively short or long, which depends on whether the email is an original or a response. But identifying the sentiment words which decide the sentiment polarity based on individual feature and becomes difficult in context to emails. The "replies and forward" communication describe the topic orientation of email messages (Bogawar & Bhoyar, 2012). Srinivasarao and Sharaff (2021a,2021b) used lexicon based approach to classify the sentiment emails. By using clustering, a Multi-Label email Classification algorithm has been proposed (Sharaff & Nagwani, 2020). Here the similarity in the text of the email attribute is used to create the groups related to the similar emails. Due to its richness in noise and unstructured format, the earlier studies on email sentiment analysis face some difficulties in the direct application of traditional sentiment analysis problems. The earlier studies advise the suitability to conduct the sentiment analysis of a document with the email dataset. Its meta-information, like sender and subject, will have the required information related to the opinion holder and entity (Shen, Brdiczka, & Liu, 2013). It is essential to find out a new methodology to discover the sequence in the email-thread and perform sentiment analysis in the documents and email data. Srinivasarao and Sharaff (2021a), (2021b) used fuzzy-model-based Gaussian clustering

* Corresponding author.
E-mail addresses: usrinivasarao.phd2018.cs@nitrr.ac.in (U. Srinivasarao), asharaff.cs@nitrr.ac.in (A. Sharaff).

https://doi.org/10.1016/j.eswa.2021.116475
Received 18 March 2020; Received in revised form 7 December 2021; Accepted 26 December 2021
Available online 3 January 2022

algorithm to identify sentiment email patterns.

Email management becomes a difficult task due to enormous data of emails and also time-consuming. So, managing the email is an important task which should be resolved. This can be achieved through the email thread generation as well as identifying the sentiments of email threads. An email thread can be defined as a grouping of message or sequence of emails for a specific user, which the user replies and forwards. The similarity between the email attributes can be used to detect the email threads. The main objective of this work is to identify the sentiment in email belonging to the same thread by using the simplified text clustering approach and identifying a sequence of threads by using a sentiment clustering method. Identification of a sequence of threads within documents in sentiment clustering depends upon on how the feature words are extracted. In a document weighting of features, frequency of words is used as features in a feature-based document sentiment analysis.

In addition to introducing the email data as input to sentimental analysis, another essential part of this study is proposing a methodology based on clustering technique to identify the sequence in the email threads to improve the sentiment analysis results at the document-level. The critical point in identifying the document's sequence will be levied on the procedure adopted in extracting the feature words. The important point in identifying the sequence in the document will be levied on the procedure adopted in extracting the feature words. Weighting or frequency of features related to the document is extracted based on Feature-based document sentiment classification. In the following example, two review fragments convey the sentiment that is a broadly positive view, but they are not the same. The rules related to the Conventional classification of document sentiment will, in general, treat the feature vector in a static way that does not consider the interaction among the documents. Both the documents that are analyzed as positive may have different sentiments depending upon the document's features position. It is explained in the following example: for the first review = positive → positive → negative → positive → negative; for the second review = positive, → positive → negative → positive. On the other hand, the new approach that performs the sequence in thread-based email sentiment analysis will introduce the sequence in a thread within sentiment analysis documents with the consideration of chronological features.

This paper aims at identifying the sequence in thread-based email sentiment analysis and also the thread sequence within the emails considering the sequential presence of features by using the Probabilistic Latent Semantic Analysis (PLSA) clustering algorithm. The main contributions of the study are described as follows:

- The sentiment analysis task is solved where the email thread sentiment is identified first, and then the sequence of the sentiment threads in the email is identified using the PLSA clustering algorithm.
- The sentiment and the thread sequence within the emails have been discovered as clustering sentiment polarity and temporal categories with the help of PLSA clusters.
- To visualize the thread's sentiment sequence by aligning it with actual email data that are characterized in sentiment features along with email data both in temporal and topic distribution.

The paper is arranged as follows. Related work is explained in Section 2. Problem statement has been described in Section 3. The clustering Approach has been described in Section 4. The proposed methodology has been detailed in Section 5. Section 6 represents the results and discussions. And the conclusions are presented in Section 7.

2. Related work

Sentiment analysis (SA) can also be referred to as opinion mining where the opinions, emotions, and attitudes are identified and extracted from the text, Zhao, Liu & Xu (2016). Mao and Lebanon (2007) used for the first time, the concept of the local sentiment using modified Conditional Random Field (CRF) to analyze the emotion within a document. Tang et al. (2015) performed the document sentiment analysis using deep learning technique which is based on word embedding. Sharaff and Soni (2018) have classified the reviews as positive and negative using the features of the products. To assess the applicability of the proposed technique, amazon data was considered for identifying the positive and negative sentiments of the product. Liu and Lee (2018) have discovered sentiment sequence of emails by using trajectory clustering algorithm and SentiWordNet (SWN) lexicon. Bespalov, Qi, Bai, and Shokoufandeh (2012) used N-gram features based supervised sequence embedding algorithm for sentiment classification. Wei and Chang (2007) identified the event episodes, and their temporal relationships are identified by using an event-evolution pattern discovery technique. Feng, Wang, Yu, Gao, and Wong (2011) proposed sequence modelling-based neural network to Performa document-level SA. An unsupervised machine learning algorithm, i.e. Probabilistic Latent Semantic Analysis (PLSA) text clustering, has been successfully applied to retrieve the information for various purposes, like common emotions detection from blogs. Liu, Lee, and Lee (2020) developed a new neural network-based methodology to classify the sentiment of multi-topics. It uses Bidirectional LSTM (BiLSTM) with topic weighting vectors and topic embeddings. Ezpeleta, Velez de Mendizabal, Hidalgo, and Zurutuza (2020) It is a proposed method for validating hypotheses to identify the email messages solved by personality recognition methods and performing sentiment analysis.

(Xie, Ge, Hu, Xie, & Jiang, 2019) used PLSA in extracting the seed emotion words in Wikipedia. Based on the seed emotion words, the features were extracted and given as input to train the model. Ezpeleta, Garitano, Zurutuza, and Hidalgo (2017) have analyzed the Short Message Services (SMS) using a combined sentiment analysis and personality recognition methods. Socher, Pennington, Huang, Ng, and Manning (2011) developed a methodology using a semi-supervised machine learning technique, i.e. Recursive Auto-Encoders (RAE) which can predict the multi-dimensional distributions related to the underlying feeling. Zhang, Ghosh, Dekhil, Hsu, and Liu (2011) proposed a hybrid technique to perform opinion mining. First, an augmented lexicon-based method has been applied for Twitter data for performing an entity-level SA. Vashishtha and Susan (2019) introduced a fuzzy based methodology to perform sentiment analysis task over the datasets of the social media posts like Twitter data.

Further, a binary classifier is trained depending upon the results received in the previous step for assigning the sentiment polarities to the corresponding tweets. Cui, Mittal & Datar (2006) performed the binary sentiment classification task, to assess the effectiveness of classifiers like Support Vector Machine (SVM), by using the input features as high order n-gram. Severyn and Moschitti (2015) performed a Deep Convolutional Neural Network-based sentiment classification by using the word embedding, which was initialized through an unsupervised neural language model. Ren, Wang, and Ji (2016) used Latent Dirichlet Allocation (LDA) for obtaining the topic distribution related to each sentence in the set of data and then RAE for learning the topic-enhanced word embeddings. For further improvement and performance, traditional approaches like Logistic Regression (LR) and SVM should be integrated. Mesnil et al. (2014) applied ensemble approaches involving both generative models like Recurrent Neural Networks (RNN and Naïve Bayes) and discriminative models for sentimental analysis. The log probabilities of the corresponding models are associated with linear insertion for extracting the sentiment and perform better compared to all competitive models. Wang, Zhang, Sun, Yang, and Larson (2015) suggested a new Part-of-Speech and Random Subspace technique (POS-RS) for SA which uses the POS analysis with the help of both function lexicon and content lexicon subspace rates for controlling the diversity related to the base learners. Heerschop et al. (2011) analyzed how the knowledge should be extracted from structural features of a document and how to enhance the performance of sentiment analysis model. Besides, the hypothesis test was conducted based on significant document

segment identification which was used for detection of sentiment and providing the score to documents. Rezaeinia, Rahmani, Ghodsi, and Veisi (2019) introduced a new methodology named Improved Word Vectors (IWV). It is proposed to increase the accuracy in the sentiment analysis related to the pre-trained word embeddings. (Ali et al., 2019) proposed a novel fuzzy ontology-based lexicon technique to perform sentiment classification. Here six more different lexicons are used for enhancing the accuracy of the pre-trained word embedding model.

(Qiu et al., 2010) extracted the opinion sentences that are related to negative sentiment to find out the sentiment topics with the help of a rule-based method which combines a sentiment lexicon and syntactic parsing. This approach has been tested in contextual advertising, i.e. the problem of linking the advertisements and Web page. Saif, He, Fernandez, and Alani (2016) proposed a lexicon-based method known as Senti Circles is discovering the latent semantics from the co-occurrence of patterns. The designed model can be utilized for updating the sentiment orientation related to the words. Carrillo-de-Albornoz and Plaza (2013) used natural language processing (NLP) for identifying the linguistic features like negation, modalities and intensifiers, and lexicons for determining the overall sentiment. Understanding the expression of people's opinion is difficult for users as well as system. Further, transforming the data as meaningful information will be helpful to control and manage the traffic and transportation services. It is essential to filter out inappropriate information in the input data for further identifying the topics and features. Valdivia, Luzón, Cambria, and Herrera (2018) used a neutrality proximity function for filtering out the neutral information before it is used for binary classification. Naseem, Razzak, Musial, and Imran (2020) proposed DICET, which is a transformer-based technique to identify the sentiment. It will encode the representation from a transformer and apply the deep, intelligent contextual embedding for enhancing the quality of tweets, and the noise is also removed here. Daudert (2021) performed financial sentiment analysis using neural network. Here, relationship and text features based on entity, word and temporal information are used. Asani, Vahdat-Nejad, and Sadri (2021) have performed study personalized system extracts user preferences with the analysis on the user opinions and further the acquired list is refined using sentimental analysis. Xie, Lin, Lin, Wang, and Yu (2021) analyzed a technique for predicting the multi-dimensional sentiment score. Basiri, Nemati, Abdar, Cambria, and Acharya (2021) have performed study for sentimental analysis using deep learning model i.e. an attention-based bidirectional CNN-RNN.

(Alkhereyf & Rambow, 2020) Proposed a method Long Short-Term Memories (LSTMs) networks used to identify the personal and Business email thread structures. Wu and Oard (2005) presented an automatic subject line emails should belong to the same conversation thread after removing any sequence of 'reply', 'fwd' messages. Sharaff and Nagwani (2016) proposed a methodology for identifying the email threads based on the threading features like people and subject. As these methods are unable to find the direct relation in messages, the other methods are intended to frame a tree structure between emails which are dependent on the reply relations. Balali, Faili, Asadpour, and Dehghani (2013) analyzed the thread structure generation based on the comments available in online fora, chats etc. Probabilistic sequence labelling model and graph-theoretic modelling are applied on asynchronous conversations. Dehghani, Asadpour, and Shakery (2012) presented a genetic programming-based methodology to reconstruct email threads. Here, the thread reconstruction is analyzed by considering it as an optimization problem. Joshi, Contractor, Ng, Deshpande, and Hampp (2011) performed email grouping to identify the concepts by creating a group of emails based on a near-duplicate detection technique. Nenkova & Bagga (2004) used an extractive summarizer-based system for generating the summary of email thread which is possible in an archived discussion. Dehghani, Shakery, Asadpour, and Koushkestani (2013) used the linear and tree structures of threads belonging to the email data, two learning approaches, i.e.LExTreC and LExLinC, are presented to reconstruct. Balali, Faili, and Asadpour (2014) analyzed an SVM-based model for automatic reconstruction of the thread structure. To train the SVM classifier for identifying the thread structure textual comments using non-textual and textual features. Nagwani and Sharaff (2017) used Non-negative Matrix Factorization (NMF) clustering along with SVM classification technique to identify the SMS threads. Here, the semantic analysis is combined with a social network to develop a new technique for identifying the top performers based on email communication (Wen, Gloor, FronzettiColladon, Tickoo, & Joshi, 2020). Nascimento and De Carvalho (2011) developed a graph partitioning algorithm related to spectral clustering. It is derived from the Laplacian matrix of a network. This is often mathematically termed as a graph. Based on particular properties, the spectral clustering technique will divide any dataset into smaller clusters. The datasets included in the same group will be having more significant similarities when compared to the datasets between the clusters. Zhou, Ye, Plant, and Böhm (2017) proposed a Gaussian Mixture Models (GMM). In recent times this technique has been popular due to its excellent performance and simplicity in its implementation. AlMahmoud et al. (2020) combined a fuzzy merging technique and an improved version of the Bond Energy Algorithm (BEA) for solving the problem of Arabic text document clustering.

3. Problem statement

As described in the above section, this paper proposes the sequence email thread and further performs the sentimental analysis. The crucial issue that must be solved in this study is finding out the sequence of the sentiment in the email thread in the document and is followed by assigning the polarity of the sentiment to PLSA clusters. The PLSA concept and its related feature vector related to the latent topics have been used to attain this. To attain this, the PLSA concept along with its related feature vector related to the latent topics has been used.

Unlike the conventional way of solving the sentimental analysis, this study solves the problem using the sequence in the thread and then solving it as an email sentiment clustering problem. This study can be divided into two main parts:

1. Find out the sequence of the sentiment in the thread of emails and identify the thread.
2. Making the sentiment polarity as clusters.

This work is mainly focused on solving one problem that aims to justify the assumption of a sequence of the sentiment in the email thread will influence the polarity of the sentiment that has to be classified or clustered. The challenges that are related to the two problems reflect two aspects. First is the loss of information while performing the transformation process and the next is an issue related to the unbalance in the email data feature distribution. The spatial information is used to extract the features, and then the documents are changed into latent topics, which are to be computed with the help of the PLSA technique. The PLSA algorithm is modified so that it will be able to assign the polarity in the sentiment at each instant that may be validated. Even though a quantitative assessment is a bit harder to undertake because of the unavailability of labeled datasets related to email, a good essence of validation to the proposed method has been given by manually prepared labeled small set of email data and comparatively larger review data that is pre-rated. Initially, the proposed method's validity has been tested on a manually prepared labeled small set of email data and is further extended over larger data. As there is no availability of a pre-labeled email dataset in the public domain, the proposed method has been evaluated over larger review data that is pre-rated for demonstrating the scalability and efficacy of the proposed work. In this work, three classes have been considered, i.e., neutral, positive, and negative, applied as an evaluation matrix. Because of the implicitness of the email data, it contains lesser emotions instead of binary classification. Both quantitative and qualitative results have been depicted depending upon the patterns identified with the justification of email messages in feature

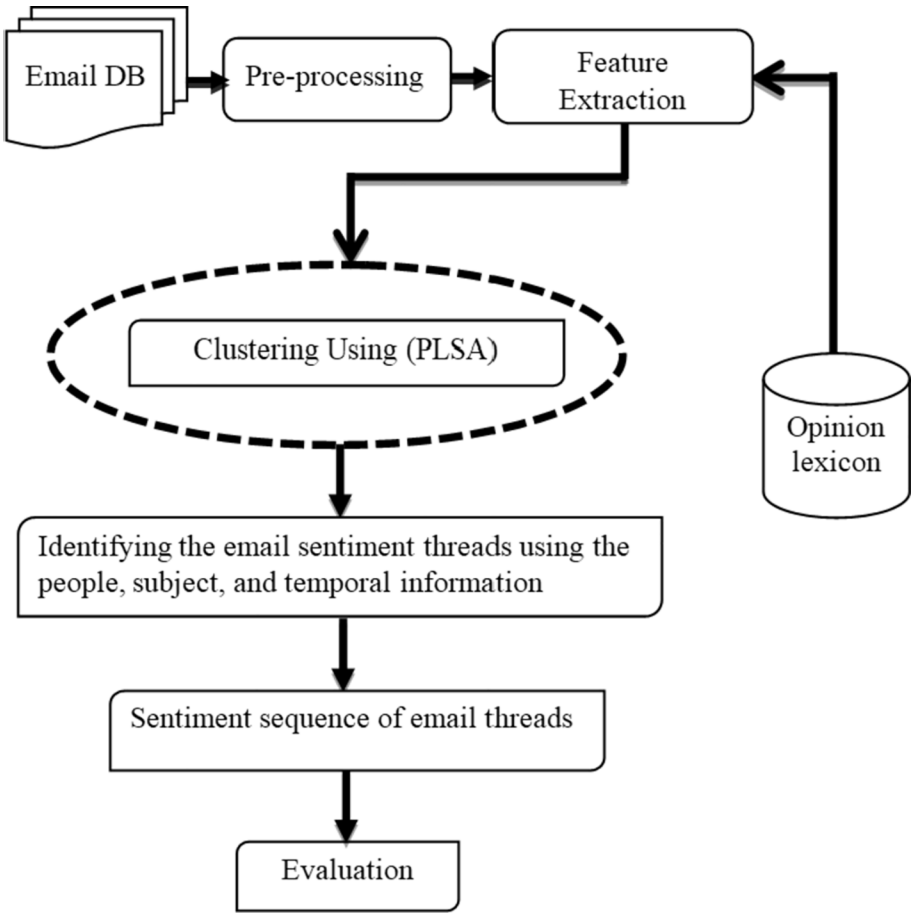

Fig. 1. Sentiment sequence of thread identification.

representations. The publicly available Enron email dataset is used in the proposed study. The PLSA based latent topics representatives are proposed, and the corresponding results of the sentiment clustering are predicted by identifying the sequence in the email thread.

4. Clustering approach

PLSA algorithm has been applied over the email dataset, which transforms the problem into probability-based representations with temporal and topic features. The original PLSA algorithm has been modified to store the topic and temporal data related to the email dataset. The original PLSA algorithm is designed to deal with the latent topic model in datasets along with more attributes, including latent topic features such as size and id only. It would be better if the sentiment polarity and the sentiment sequence of email threads, as the PLSA algorithm, will result in clusters representing the probability likelihood parameter estimation values.

Clustering is related to the grouping of similar emails, in one group and isolating them from the other group which incorporates information on a different topic. Grouping emails into various clusters facilitates maintenance. PLSA algorithm reveals the likelihood of email belongs to each cluster in terms of email attributes including to-address, from-address, content, subject, CC-list, BCC-list, and attachments. To effectively group similar emails, the proposed work employs PLSA text clustering technique to create email clusters.

4.1. Probabilistic latent semantic analysis

Hofmann (2013) proposed PLSA, which is a probabilistic model where a word in a document is featured as an example of a mixture model. The collaborative components are multinomial distributions which are independent conditionally.

Consider a document collection $c \in C = \{c_1, c_2, \cdots c_n\}$. By using the PLSA method as clustering algorithm, first a latent topic model is defined which is associated with an unobserved latent variable $t \in T = \{t_1, t_2, \cdots t_k\}$ corresponding to the word $s \in S = \{s_1, s_2, \cdots s_m\}$ in $c \in C$. The three steps involved in the model construction are:

i) Selecting a document (c) which has a probability $P(c)$;
ii) Choosing a latent topic (t) which has a probability $P(t|c)$;
iii) Generate a word s which has a probability $P(s|t)$.

Assume each pair (s, c) is independently generated the document c and word s are conditionally independent. For a given latent topic t, the joint probability $P(s, c)$ is computed as given in Eq. (1).

$$P(s,c) = P(c)\sum_{t\in T} P(t|c)P(s|t) \tag{1}$$

On the basis of Bayes' rule, the above equation is rewritten as Eq. (2).

$$P(s,c) = \sum_{t\in T} P(c|t)P(s|t)P(t) \tag{2}$$

By increasing the word log-likelihood function, the best parameters $P(c|t)$, $P(s, t)$ and $P(t)$ should be determined by Eq. (3).

$$L = \sum_{c\in C}\sum_{s\in S} n(s,c)\log P(s,c) \tag{3}$$

where $n(c, s)$ represents how often the word s occurs in the document c. Niu and Shi (2010) proposed the classical Expectation Maximization (EM) algorithm which has been used for performing maximum

likelihood parameter estimation. This infers the unknown conditional probability values distribution in PLSA. During the Expectation (E) and Maximization (M) steps conditional probabilities are iteratively estimated as fix points. The E and M steps are performed till the convergence is achieved.

A. **Expectation step (E):** In this step the posterior probabilities for the latent variables are computed and can be represented using the given Eq. (4).

$$P(t|c,s) = \frac{P(t|c)P(s|t)}{\sum_{t\in T} P(t|c)P(s|t)} \tag{4}$$

where p is a stranded parameter that is the previous iteration of the EM algorithm estimates.

B. **Maximization step (M):** Here, probability distributions are updated using the posterior probabilities determined in Eq. (5)-(6).

$$P(s|t) = \frac{\sum_c n(c,s)P(t|c,s)}{\sum_s \sum_c n(c,s)P(t|c,s)} \tag{5}$$

$$P(t|c) = \frac{\sum_s n(c,s)P(t|c,s)}{\sum_t \sum_s n(c,s)P(t|c,s)} \tag{6}$$

Here $n(c,s)$ is the number of times the word s in a document c. The conditional probability $P(t|c)$ of topic t on given document c, can be calculated from the Bayes rule eq. (7).

$$P(t|c) = \frac{P(c|t)P(t)}{P(c)} \tag{7}$$

By using word s, document c, andlatent topic t, with the optimal parameters $(P(s|t), P(c|t), P(t).)$ which are estimated using the EM algorithm. Based on these steps the clusters are generated. The clustering is performed by assigning the label t^1, that is having maximal value of $P(t|c)$ as the maximum possible cluster label to c. For all the cluster labels, $P(c)$ is constant in Eq. (8).

$$t^1 = argmax_{t\in T} \frac{P(c|t)P(t)}{P(c)} = argmax_{t\in T} P(c|t)P(t) \tag{8}$$

From an unlabeled data, by using probability distributions over fixed topics the PLSA is used to estimate the topic-term, and document topic distributions.

5. Proposed methodology

This section presents the proposed methodology to identify the sentiment flow in the documents and has used sentiment polarity clustering as three classes: neutral, positive, and negative, using the PLSA algorithm. The proposed work combines topic modeling and text mining, which uses the PLSA clustering algorithm.

The proposed technique based on PLSA text clustering algorithm is used to determine the sentiment flow in the documents by clustering the sentiment polarity into three groups which include positive, negative and neutral. Further, the sentiment in thread identification is also performed.Fig. 1 represents the flowchart of the proposed technique for identifying the sentiment sequence and the thread information within the documents. Computational steps algorithm for the proposed method is depicted in algorithm 1. Initially, the email dataset is collected to analyze, and then textual pre-processing has been performed. This pre-processing step includes four important techniques, which are stopping, tokenization, PoS tagging, and stemming. Consequently, the proposed work uses LDA topic selection, TF-IDF document matrix, bag of words (BOW) model and SentiWordNet 3.0 words for feature extraction. These are explained in the below sub-sections.

5.1. Pre-processing

The Natural Language Processing techniques, like tokenization, stemming, and sentiment analysis cleaning steps, like Part-of-Speech tagging and stop words removal, are employed in the proposed study for pre-processing. Pseudocode 1. depicts the pseudocode for the pre-processing algorithm.

- **Tokenization**
  The large paragraphs are said to be chunks of text and split into tokens which are sentences. Further, these may be divided into words. Let us consider the sentence, “I like apples and bananas commonly” and after performing the tokenization it is broken as I, like, apples, and, bananas, commonly.
- **Stemming**
  Stemming is performed on the text collection. To perform perfect analysis, stemming is used, which converts the words into their root forms to uniformly present all the words in the text collection. For instance, the words ‘cooks’, ‘cooking’ and ‘cooked’ are converted to the word ‘cook’, which is the base to make sure the concept is unique for textual analysis.
- **Stop Words Removal**
  In a text, we find much word which does not have a specific sense when considered in pattern analysis. e.g. The, are, is, of, was and so on are not useful in knowledge discovery; such kind of words can be termed as stop/useless words. These unnecessary words will be eliminated from the text. This is performed using a list of words, which contains all the useless (irrelevant) words which should be taken out from the text corpus to prepare the dictionary of relevant or useful words as a text for better analysis.
- **Part-of-speech (POS)**
  POS tagging will assign a word belonging to its grammatical category, to understand its significance in the sentence. Traditional parts of speech are “verbs”, “nouns”, “conjunctions”, “adverbs”, etc. POS taggers use a sequence of words as input and give a list of tuples as output. Here, each word is connected with the related tags.

Algorithm 1. Computational steps for the proposed algorithm

Step1: Data acquisition & pre-processing
  1.1. Acquisition:- Acquire data from the database which contains email data
  1.2. Pre-processing:- Perform tokenization, stopping, stemming and POS tagging to get a
  Qualitative and quantitative data.
Step 2: Feature Extraction:
  2.1. Key word search:-Based on database choose relevant topics from LDA model to Extract topics.
  2.2. Bag of words:- Find the repeated words which are available in a BOW and assign they count.
  2.3. TF-IDF:- Using TF-IDF, formulated BOW as a matrix known as Term email Matrix.
  2.4. SWN lexicon:- Extracting opinion words.
Step 3: Clustering:- Formulated positive, negative, neutral clusters using PLSA clustering algorithm
Step 4: Identify the sentiment threads in each cluster based on reply and forward of an email
Step 5: Describing the sequence of sentiment threads in each cluster
Step 6: Extract the threads based on similarity such as people, subject & time.
Step 7: Evaluate sentiment thread clusters using performance parameters

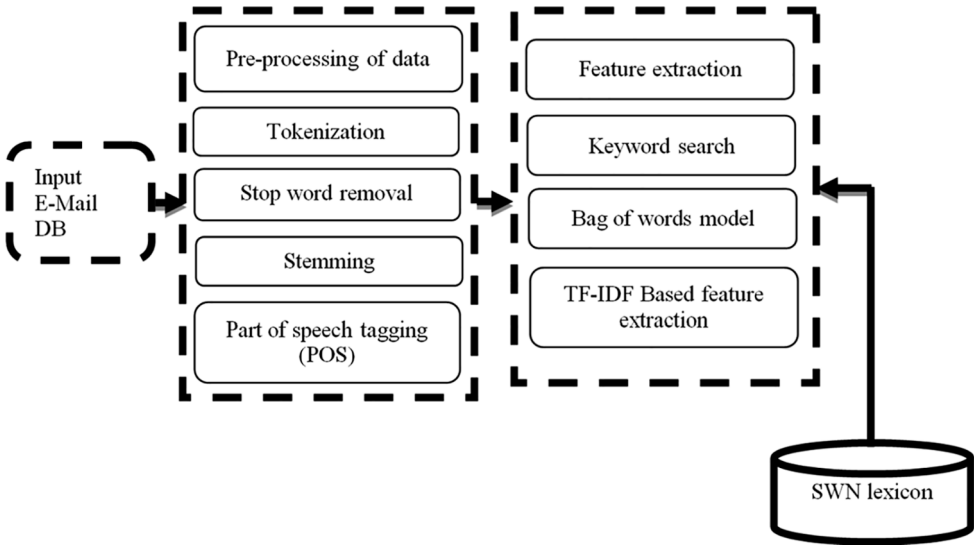

Fig. 2. Preprocessing and Feature extraction phase.

Pseudo code 1. Email data pre-processing

1: Input: Defined E as the entire email set contain email message $e_1, e_2, e_3 \cdots .. e_n$, w as a set of tokens in each email messages contain words list $w_1, w_2, w_3, \cdots w_i$.
2: Output: Each email message $e_n \in E$ represented with a collection of tokens w_i in is fined word list W.
3: for each E-mail messages $e_n \in E$ do
4:Tokenize e_n using tokenize () function;
5: for each token $w_i \in W$ do
6: if $w_i \in$ English stop word list then
7: Remove w_i from W;
8: Stemming w_i using stem () function;
9: Generate POS tagging w_i using tagging () function;
10: Return each email e_n with a list of refined words T;
11: end if
12: end for
13: end for

5.2. Feature extraction

The feature extraction procedure is divided as feature topic selection, detection and extraction. Medhat, Hassan, and Korashy (2014) used the features that are used in most sentiment analysis studies and comprised of n-gram, frequency, negation, and POS tags. He and Zhou (2011) identified some feature extraction techniques were domain-specific for a particular type of sentiment analysis. Joty, Carenini, and Lin (2011) examined the strengths and weaknesses of the methods previously proposed. Mohey (2016) explained that the research study on email sentiment analysis is limited, so a suitable feature extraction method for email data requires further studies. Consequently, the proposed work uses LDA topic selection, TF-IDF document matrix, bag of words (BOW) model and SentiWordNet 3.0 words for feature extraction.

5.2.1. Topic selection

In the information retrieval (IR) system, feature extraction is one of the stages which extract the unique feature values related to textual email messages. Among the various available feature extraction methods, Latent Dirichlet Allocation (LDA) is one to derive topic keywords as the order of the keywords will directly affect the semantics. The order of topic keywords generated using LDA is useful to understand the semantics of a topic. For example, when LDA is applied to an Enron corpus, common words like Valuation, Company, Enron, Business, etc., are most relevant to the topics generated. All pre-processing and feature extraction steps are as shown in Fig. 2.

5.2.2. Bag of words (BOW) model

It is a representation scheme commonly used in NLP and Information Retrieval (IR). Ma, Sun, Wang, and Lin (2018) proposed a multiset of the words in which the text in a document is denoted as the bag, keeping multiplicity and ignoring the grammar and word order. El-Din (2016) BOW model represents the text which explains the occurrence of words within a document. After the keyword searching BOW model is applied to count how many times each word has appeared in a document, this model is a secure method to understand the text, which is used for retrieving information for the text corpus by ignoring the sentence structure. A sample BOW model is present in Fig. 3.

5.2.3. TF-IDF based feature extraction

TF-IDF is a very well-known technique in the field of NLP and is used in the execution of the proposed algorithm (Uğuz, 2011). Determination

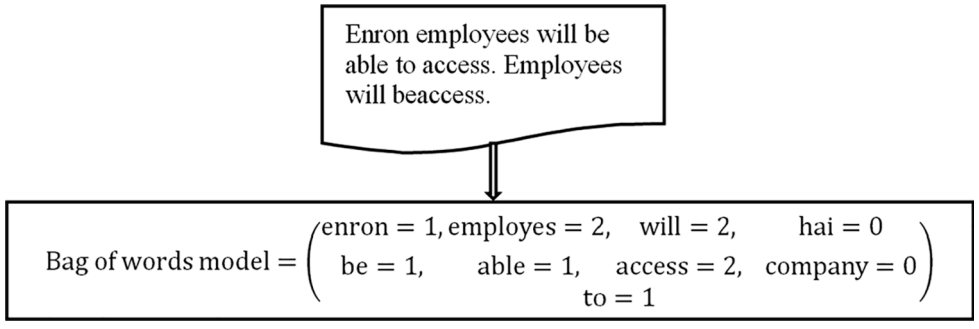

Fig. 3. Sample bag of word model.

of TF-IDF will be done by the relative frequency of words using the inverse proportion of words in the entire document text. Two elements are used to determine the value: TF -term frequency and IDF - inverse document frequency. Eq. (9) shows the classical formula of TF-IDF which is used to weight the term.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{9}$$

where $w_{i,j}$ represents the weight of the i th term in j themail, N represents the number of emails, $tf_{i,j}$ represents the term frequency of i th term in j themail and df_i represents the email frequency of i th term in the collection. In this work, TF-IDF is treated as Term Frequency and Inverse email Frequency (TF-IEF) which generates the email matrix (email frequency) using the bag of word model as input. Here, each word can be represented by matrix representation of entire email words. The email frequency of each expression is counted in the matrix.

5.2.4. SentiWordNet (SWN)

(Baccianella, Esuli & Sebastiani, 2010) used a refined and structured English sentiment lexicon called as SWN for generating the initial sentiment feature. SWN has been derived from WordNet, which is a lexical reference that is publicly available. Pang et al. (2002) used SWN, which is broadly used in opinion extraction and sentiment classification tasks. According to recent research articles, it is also used to extract the feature in sentiment analysis related works. Tai, Tan, Lin, and Chang (2015) used LDA along with SentiWordNet to detect mental disorder using the emotion scores calculated from the user's diaries. SWN consists of 147,305 sentiment phases with six attributes which can uniquely identify each item. Each sentiment phase can be identified using combined POS and gloss tags. A positive and negative score that is generated using a frequency-weighted average of its relevant cognitive synonyms using a semi-supervised learning technique is also used to identify each sentiment phase. In each step, sentiment has been assigned a value using an overall objective score. Consider $S = \{s_1, s_2, \cdots\cdots s_i\}$ as a set of sentiment phases and $V = \{v_1, v_2, \cdots\cdots v_i\}$ as a set of sentiment values in SWN. The new sentiment value of each sentiment phase v_i can be computed using Eq. (10):

$$V_i = 1 - (Posvalue_i + Negvalue_i) \tag{10}$$

Finally, sentiment scores are generated using word net lexicon. The documents should be transformed into a vector space using the temporal position of each feature, and the corresponding values are generated based on SWN lexicon.

5.3. PLSA based sentiment email clustering

Probabilistic Latent Semantic Analysis (PLSA) is one of the better algorithms for modeling document collections. The probability matrices of corpus set related to word topic and document topic are computed with the help of PLSA based method. On the basis of dispersal of words related to a topic, first PLSA will cluster the words and next will merge likely topics, further will find out the category of emotion on the basis of sentiment SWN library. Moreover, for realizing the sentiment email clusters document-topic probability distribution will be used as benchmark. When applied to the document clustering, latent topics corresponding to PLSA will be identified over the clusters. They may be very restrictive in some scenarios, which have more issues when compared to document clusters.

In this paper, PLSA text clustering algorithm is used for generating email clusters. Email clusters are created using PLSA algorithm using Eq. (8). PLSA is an algorithm which is based on topic modelling i.e. clusters can be grouped by content similarity and topics occurrence of data. The email content similarity calculation is obtained according to Eq. (11). In the equation $w_{k,i}$ and $w_{k,j}$ denote the k th word in emails M_i and M_j respectively. Based on the content similarity by using SWN sentiment email messages are clustered into a group of messages.

Table 1
Confusion matrix.($K = 3$)

		Predicted		
		Neutral (PNe)	Positive (PP)	Negative (PN)
Actual	Neutral (ANe)	(ANePNe)	(ANe PP)	(ANe PN)
	Positive (AP)	(AP PNe)	(AP PP)	(AP PN)
	Negative (AN)	(AN PNe)	(AN PP)	(AN PN)

$$Sim_{content}(M_i, M_j) = \frac{\sum_{K=1}^{n} w_{k,i}.w_{k,j}}{\sqrt{\sum_{K=1}^{n}(w_{k,i})^2}.\sqrt{\sum_{K=1}^{n}(w_{k,j})^2}} \tag{11}$$

5.4. Clustering using sentiment in threading features

(Cselle, Albrecht, & Wattenhofer, 2007) considered many parameters in the study to recognize the similarity of the email. The parameters are Sender rank, Sender percentage, Sender answers, Time, Known people, Reference count, People count, Known references, Cluster size and has attachment. Among these parameters, five parameters namely Sender rank (subject similarities), Sender percentage, Time, People count (People similarities), and Cluster size are used for identifying the sentiment in an email thread and also for thread identification from the email clusters. The subject of an email is compared with the other emails subject. Thus, provides subject similarity based threads. When the receiving time between the pair of emails is less than λ (time threshold), then the pair can participate in the same thread on the basis of people similarity, subject similarity and time similarity. Time similarity is one of the best among all other similarities because it can identify a maximum number of threads. Also based on subject, a people similarity between pairs of emails is computed using Eqs. (12)–(14).

$$sim_{subject}(m_i, m_j) = \frac{2|s_i \cap s_j|}{2|s_i| + |s_j|} \tag{12}$$

$$sim_{people}(m_i, m_j) = \frac{2|people(m_i) \cap people(m_j)|}{|people(m_i)| + |people(m_j)|} \tag{13}$$

$$sim_{Time}(t_i, t_j) = \frac{2|Time(m_i) \cap Time(m_j)|}{|Time(m_i)| + |Time(m_j)|} \tag{14}$$

5.5. Performance measurement

To evaluate the sentiment in threads and email threads that are generated using the proposed method, it has been converted into a classification problem. After categorizing the problem at three levels, the performance is assessed through different classification parameters like F-measure, precision, recall and accuracy.

Accuracy – the accuracy is the ratio of correctly classified email messages participating in threads to the actual number of email messages. The accuracy computed as in Eq. (15).

$$Accuracy = \frac{\text{Number of messages correctly classified}}{\text{Actual number of messages}} \tag{15}$$

The F-measure is the combination of precision and recall. Precision (P) is defined as the ratio between the actual neutral prediction neutral and all the neutral (i.e. sum of actual neutral prediction neutral and actual positive prediction neutral and actual negative prediction neutral). While recall (R) is defined as the ratio between the actual neutral prediction neutral and actual neutral prediction neutral with actual neutral prediction positive and actual neutral prediction negative. The corresponding equations of F-measure, precision and recall are given in Eqs. (16)–(18). Accuracy may also be calculated based on the confusion matrix as shown in Table 1. Here K represents the number of

Table 2
Distribution of the thread size clustering information of sentiment.

ThreadSize	PLSA			K-Means			Spectral			Gaussian Mixture Model (GMM)			Fuzzy Merging Clustering		
	Positive (P) Threads	Negative (N) Threads	Neutral (Ne) Threads	Positive (P) Threads	Negative (N) Threads	Neutral (Ne) Threads	Positive (P) Threads	Negative (N) Threads	Neutral (Ne) Threads	Positive (P) Threads	Negative (N) Threads	Neutral (Ne) Threads	Positive (P) Threads	Negative (N) Threads	Neutral (Ne) Threads
2	7501	7730	7307	3514	4497	4015	3145	7538	7500	5892	5904	4895	6742	6912	6579
3	2657	2636	2940	4004	4862	5933	3517	2736	2774	2422	2454	3281	2387	2674	2447
4	2302	2459	2329	381	398	224	217	2361	2261	3170	2119	1548	2178	2272	3078
5	848	849	760	1147	1247	1387	1135	822	892	1873	1722	1637	849	829	897
6	587	636	522	681	879	959	571	573	627	601	648	755	517	979	817
7	522	498	480	512	712	858	744	494	440	468	442	301	548	671	547
8	280	330	764	674	757	452	517	294	327	251	201	152	480	327	615
9	154	174	152	204	347	206	187	162	129	180	107	88	244	154	331
10	118	158	145	89	179	200	68	147	218	152	111	155	188	129	134
10–20	430	413	467	687	812	710	579	562	529	308	219	235	286	341	386
20+	68	79	37	78	97	68	73	83	45	89	95	83	49	67	49
Total	15,467	15,962	15,903	11,971	14,787	15,012	10,753	15,772	15,742	15,406	14,022	13,130	14,468	15,355	15,880

clusters ($K = 3$). It is the ratio between the true observations and all observations and is depicted in Eq. (19).

$$Precision = \frac{\text{ANePNe}}{ANePNe + APPNe + ANPNe} \tag{16}$$

$$Recall = \frac{ANePNe}{ANePNe + ANePP + ANePN} \tag{17}$$

$$F - Measure = \frac{2 \times \text{Precision} \times \text{Recall}}{Precision + Recall} \tag{18}$$

$$Accuracy = \frac{ANe\ PNe + AP\ PP + AN\ PN}{Total\ email\ messages} \tag{19}$$

6. Experimental results and discussion

For evaluating the proposed model for document clustering, a standard dataset for clustering corpora is used with the class labels as objective knowledge to reflect the implicit structure of the dataset. The following will explain the dataset that has been used along with the measures that are taken for evaluating the performance of the clustering. The main focus is on the clustering problem. The clusters are the true underlying topics that are extracted with the help of PLSA. The assessed conditional probability distribution function $P(z_k/d_i)$ may be used for inferring the cluster label corresponding to each document.

In this step, the experimental detail of the proposed work is presented. The analysis of the proposed method based on PLSA has been compared with some standard algorithms like K-Means, Spectral Clustering, Gaussian Mixture Model (GMM) and Fuzzy Merging Methods. The experiments were performed using the python programming language. The experimental studies were performed on Enron email corpus dataset, which is freely available [4]. The dataset consists of 205,120 email messages. The main idea of the research is “sentiment clustering, the sequence of messages and thread identification”. The experiment is conducted on 205,120 Enron email messages which are clustered into three clusters like positive, negative and neutral groups. Using the PLSA based clustering algorithm, it was found that 15,467 emails fall in Positive (P) threads, 15,962 in Negative (N) threads and 15903in Neutral (Ne) threads shown in Table 2. A total of 169,717 (82.74%) of the 205,120 messages are covered in the PLSA method holds all the email message for thread identification. The average thread size of each sentiment cluster is 3.79 (P), 3.50 (N) and 3.46 (Ne). The thread size can be defined as the number of emails that participate in a single email thread. Consider an example; if the thread size is three, then it shows that the thread comprises of three emails. The size of the thread is very small for the majority of the threads. The distribution of thread size and clustering information of sentiment is explained in the Tables 1 and 2.

As shown in Table 2 three sentiments email clusters have been generated using PLSA, K-Means, Spectral, Gaussian mixture model and Fuzzy Merging clustering methods, respectively. All the emails which are available in the dataset are covered by three clusters for PLSA, K-Means, Spectral and Gaussian mixture model algorithms and these cluster sizes are standard for text mining method. Table 3 describes the total email threads which are participating in email threads, Average thread size per cluster, number of threads and percentage of email in threads. The average thread size obtained from PLSA is 3.79 Positive (P), 3.50 Negative (N) and 3.46 Neutral (Ne), from K-Means is 3.13 Positive (P), 3.29 Negative (N) and 3.40 Neutral (Ne), Spectral clustering algorithm is 3.42 Positive (P), 2.64 Negative (N) and 2.90 Neutral (Ne), Gaussian mixture model is 3.46 Positive (P), 3.55 Negative (N) and 3.34 Neutral (Ne)and from Fuzzy Merging Clustering algorithm is 3.74 Positive (P), 3.41 Negative (N) and 3.36 Neutral (Ne) respectively.

6.1. Parameters for clustering algorithms

In this section the most appropriate parameters corresponding to

Table 3
Email distribution over threads.

Algorithm	Clusters	Threads per cluster	Emails in threads per cluster	Average thread size per cluster	Total emails in threads	Percentage of emails covered in threads (Percentage of 205120)
PLSA	Positive (P)	15,467	58,670	3.79	169,717	82.74
	Negative (N)	15,962	55,948	3.50		
	Neutral (Ne)	15,903	55,099	3.46		
K-Means	Positive (P)	11,971	37,524	3.13	137,357	66.96
	Negative (N)	14,787	48,781	3.29		
	Neutral (Ne)	15,012	51,052	3.40		
Spectral	Positive (P)	10,753	36,781	3.42	124,276	60.58
	Negative (N)	15,772	41,781	2.64		
	Neutral (Ne)	15,742	45,714	2.90		
GMM	Positive (P)	15,406	53,432	3.46	147,235	71.77
	Negative (N)	14,022	49,832	3.55		
	Neutral (Ne)	13,130	43,971	3.34		
Fuzzy Merging Clustering	Positive (P)	14,468	54,178	3.74	160,030	78.01
	Negative (N)	15,355	52,381	3.41		
	Neutral (Ne)	15,880	53,471	3.36		

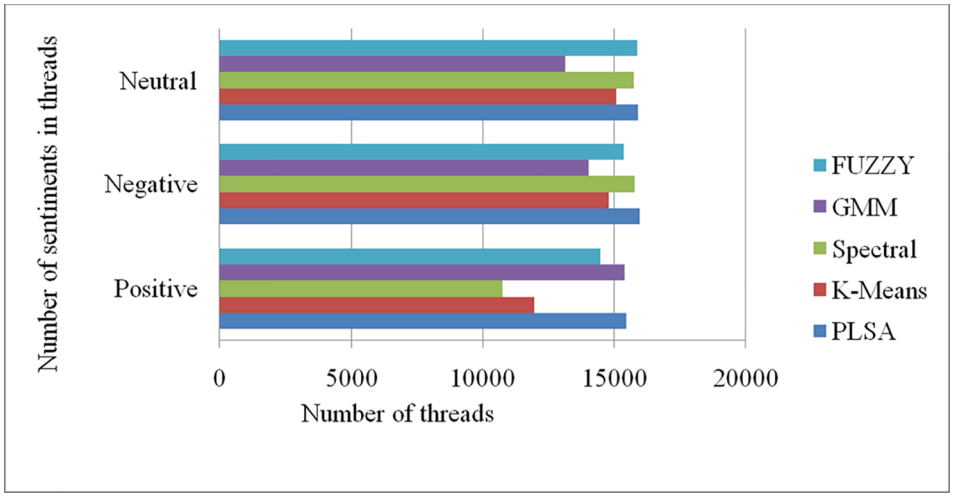

Fig. 4. Thread distribution over the clusters using various clustering algorithms.

PLSA based clustering algorithm is presented. The proposed model includes the subsequent parameters: 1) Number of topic = k, 2) Hyper parameter α, which is automatically tuned by the model and 3) Hyper parameter γ. In this work, the study has been performed to know the dependence of the number of inferred topics on the parameter K, which was varied in the range [0:001;1]. In the proposed work, by increasing the values of hyper parameters Expectation Maximization (EM) training has reached the convergence criteria after five iterations, while the other methods take about 40 iterations. At the beginning of the learning procedure, the parameter values will be equally valued because the conjugate priors smears out the conditional probabilities $P(z/u)$ and $P(r|i,z)$. EM algorithm for the Likelihood function is used to determine the above parameters. Also we compared the performance of the following methods k-means, spectral clustering, GMM and Fuzzy clustering algorithm. The exiting clustering algorithm parameters are K-Means: 1) Number of clusters = 3, 2) Maximum iterations = 30, 3) Random state = 0. Second, parameters of spectral clustering are: 1) Number of clusters = 3, 2) Hyper parameter = λ, 3) Gaussian kernel matrix = θ. Third GMM algorithm parameters are: 1) Number of clusters = 3, 2) Maximum iterations = 30, 3) Gaussian mixture component = α_i. Finely Fuzzy clustering algorithm parameters are: 1) Number of clusters = 3, 2) Membership value = μ_ij, 3) Fuzzifier hyper parameter value = m. Based on the above clustering algorithms the sentiment information is explained in the Tables 2 and 3. Also thread distribution of PLSA, K-Means, Spectral, GMM and Fuzzy Clustering algorithms is shown in Fig. 4.

The thread distribution of PLSA, K-Means, Spectral, GMM and Fuzzy Merging Clustering algorithms is shown in Fig. 4. PLSA clustering based

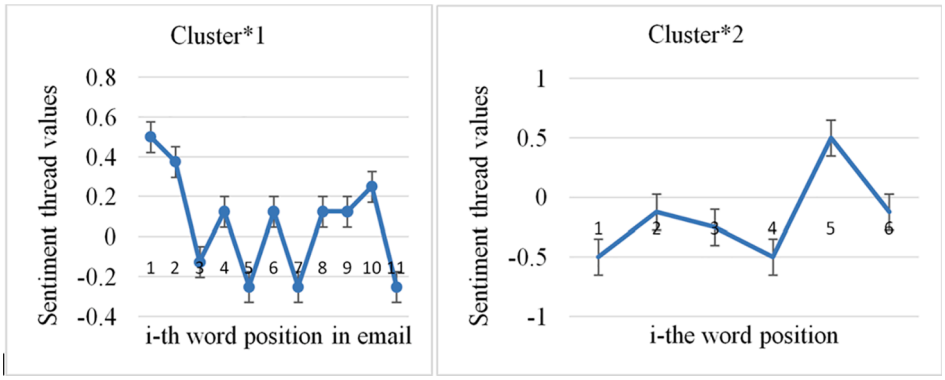

Fig. 5. Two sentiment sequences of threads identified using PLSA algorithm.

Table 4
Comparative study.

Ref	Method	Thread feature					Performance Metric			
		People	Subject	Time	All	Email Thread sentiment sequence	Accuracy	Precision	Recall	F-Measure
Dehghani et al. 2012	Single-link clustering	✔	✔	–	✔	–	–	✔	✔	–
Balali et al., 2014	Genetic programming	✔	✔	–	–	–	–	✔	✔	–
Kooti et al. 2015	SVM	–	–	–	–	–	✔	✔	✔	✔
Sharaf and Nagwani, 2016	LDA/NNMF	✔	✔	–	✔	–	✔	✔	✔	✔
Yousefpour et al., 2016	Bagging algorithm	–	✔	✔	–	–	✔	–	–	–
Liu & Lee, 2018	TRACLUS algorithm	–	–	–	–	–	✔	–	–	–
Ali et al., 2019	LDA	–	–	–	–	–	✔	✔	✔	
Proposed model		✔	✔	✔	✔	✔	✔	✔	✔	✔

'√' = achieved, '–' = not achieved.

Table 5
Sentiment in thread clustering result with email messages.

Sentiment sequence of thread	S. no.	Topic	Email Thread Message
0.5 0.375 −0.125 0.125 −0.25 0.125 −0.25 0.125 0.125 0.25 −0.25 Polarity = 0.75	1 2 	Meeting West power origin update 	{153386:[kill:-0.5,public:0.125, highest:-0.25,lowest:0.5,make:0.5, public:0.125, like:0.125,global:0.375, well:0.375,regard:0.125]} {153514: [main : 0.375, differ : 0.25, want : 0.25, date : 0.125, formal : −0.25, higher : 0.5, fit : 0.75, date : 0.125, black : −0.375, still : −0.125, better : 0.875, high : −0.25, good : 0.75, need : 0.125, talk : −0.125, contract : 0.25, risk : −0.25, green : −0.125, expect : 0.25, get : 0.125, got : 0.125, difficult : −0.75, master : 0.625.]}

on thread generation has maximum thread size as compared to spectral clustering, k-means clustering, GMM and Fuzzy merging clustering algorithm. In particular time frame only few users participate to give common information in same email (reply) which signifies very less similarity in email contents. This may be due to lack of common information in selected text. The most common information contain in small set of emails because large number of small size threads are presented. On this basis the emails will participate in a common thread.

Fig. 5 shows two sentiment sequences of threads using PLSA. It presents the sentiment values and positions of i^{th} word in the email threads. As shown in Fig. 5 both clusters are showing sequence of sentiment value in threads from negative to positive (pattern of most email threads) with sentiment polarity of 0.75 and −0.125 for cluster 1 and cluster 2 respectively. The average value of each cluster is calculated to find out the final sentiment polarity. Considering the second cluster, the final polarity value −0.125 is computed based on the summation of seven sentiment values −0.5, −0.125, −0.25, −0.5, 0.5, −0.125 and 0.125 divided by the number of sentiment values. Table 3 visualizes the details of sentiment flow in threads of each cluster.

For justifying the importance of the sequence of the sentiment of threads within a document and for improving the visualization of the sentiment sequence of the cluster, the email messages are clustered into different topics once the temporal classification is done for three clusters generated by using PLSA algorithm. The first group which is clustered as positive consists of 58,670 threads, while the second group which is clustered into negative consists of 55,948 threads and the neutral cluster, consists of 55,099 threads. To have the idea on the influence of sentiment sequence of threads, Table 5 shows the results which include variation in sentiment values within email messages from PLSA clusters, and also explained the email messages with feature words and their corresponding values of the sentiment. Fig. 5 shows the sentiment sequence of threads identified using the PLSA algorithm where the cluster1 is more fluctuated than the cluster 2 in terms of the difference in the sentiment of the thread.

Further, the proposed methodology has been evaluated by comparing different email thread related issues. These are represented in Table 4. The clustering and classification algorithms compared includes single link clustering (Balali et al., 2014; Dehghani et al., 2012) while the remaining techniques are (Kooti et al., 2016; Sharaff & Nagwani, 2016; Yousefpour, Ibrahim, Hamed, & Yokoi, 2016).

The evaluation of sentiment of thread and the thread information is performed by using accuracy, precision, recall and F-measure on the threads that are generated through the conversation of the emails between the internauts. The results are compared with different algorithms based on the thread features like subject similarity; people similarity, time similarity and three similarities and are shown in Table 6. The results indicate that maximum accuracy, precision, recall and F-measure has been achieved from the PLSA clustering algorithm for identifying thread information when compared to K-Means, Spectral GMM and Fuzzy Merging Clustering algorithm for all similarities. Table 6 indicates that PLSA based proposed methodology performs with good accuracy. By using the methodology adopted in this article, the emotional

Table 6
Performance parameters for validating thread.

Clustering Algorithm	Parameters	Email Thread Features											
		People Similarity			Subject Similarity			Time Similarity			All similarities		
		Sentiment clusters			Sentiment clusters			Sentiment clusters			Sentiment clusters		
		Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative
PLSA	Precision	88	85	85	90	86	87	93	90	91	91	91	89
	Recall	82	88	89	84	90	89	88	93	93	93	86	91
	F-Measure	85	86	87	87	88	88	90	91	92	92	88	90
	Accuracy	86			87			95			89		
K-Means	Precision	77	71	70	78	72	72	76	69	70	79	75	75
	Recall	68	75	76	69	77	78	67	74	74	73	73	78
	F-Measure	72	73	73	73	74	75	71	71	72	76	74	76
	Accuracy	73			74			71			73		
Spectral Clustering	Precision	74	67	67	75	68	69	73	65	67	75	69	70
	Recall	65	72	73	65	73	74	63	71	72	67	75	73
	F-Measure	69	69	70	70	70	71	68	68	69	71	72	71
	Accuracy	69			70			68			69		
GMM	Precision	78	70	72	78	73	73	74	70	78	76	68	74
	Recall	69	75	76	71	78	76	66	79	69	65	71	71
	F-Measure	73	72	74	74	75	74	70	74	73	70	69	72
	Accuracy	74			75			78			76		
Fuzzy Merging Clustering	Precision	81	83	83	85	88	81	87	84	86	90	89	93
	Recall	79	80	79	79	84	78	83	81	78	88	87	89
	F-Measure	80	81	81	82	86	79	85	82	82	89	88	91
	Accuracy	76			78			82			79		

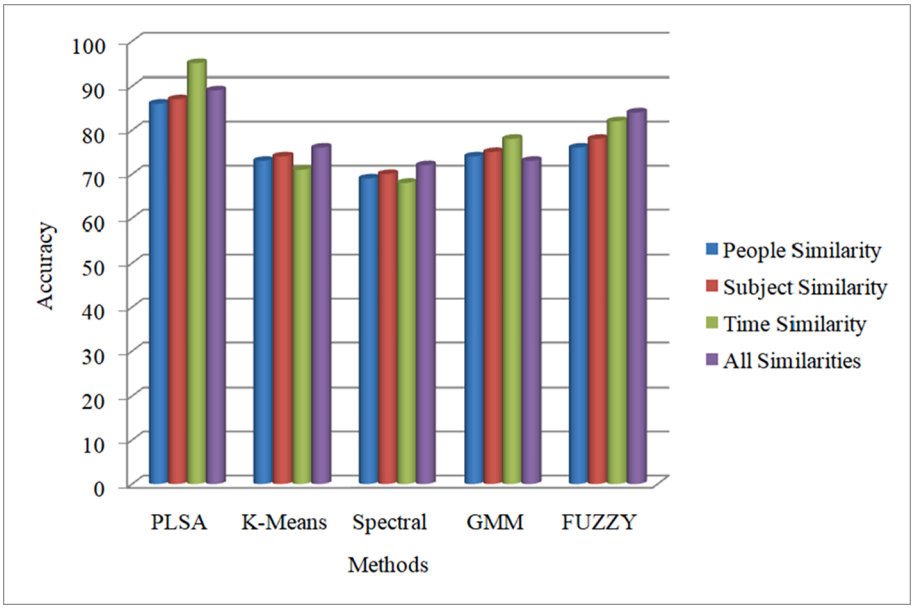

Fig. 6. Comparative performances of clustering algorithms.

Table 7
Comparison between proposed and existing methods by *t*-test validation.

Proposed Method		P-value	H-value
		4.20e-13	1
Existing Methods	K-Means Clustering	3.19e-05	1
	Spectral Clustering	1.67e-o4	1
	GMM Clustering	3.89e-45	1
	Fuzzy Merging Clustering	4.13e-1	1

distribution of people corresponding to various topics can be thoroughly studied as this is having greater importance for analyzing sequence of the sentiment in the email thread.

Fig. 6 Comparative performances of clustering algorithms with all similarity measures. The results indicate that maximum accuracy has been achieved from the PLSA clustering algorithm for identifying sentiment sequence of email thread and thread information when compared to K-Means, Spectral, GMM and Fuzzy Merging Clustering algorithm for all similarities. The statistical test, i.e. *t*-test, has been performed for evaluating the proposed method and compares it with other methods. The test results have shown statistically significant improvements by using the proposed method over the existing techniques. The numerical results attained using *t*-test is shown in Table 7. Here, the H-value represents that the null hypothesis can be rejected at the 5% significant level. When the value of H is "0″, this represents that it is not statistically significant. And when the value of H is "1", this represents that it is statistically significant. It also shows that the probability value of the proposed method is better than other methods.

7. Conclusion

A sentiment based clustering algorithm is proposed for identifying emotion in email threads, the sequence of threads and thread identification. The proposed algorithm involved in two-level clustering that provides the exact identification of sentiment in email threads and thread identification. Next, the outcome of this algorithm is compared with the standard clustering algorithms concerning the overall percentage of emails covered in threads. Thus confirms its superiority. It can also identify the thread based on people, subject and time similarity and compared with the existing clustering algorithms. Among all such similarities, the proposed algorithm with time similarity provides better accuracy, precision, recall and F-measure and for validating threads. Furthermore, a statistical test has also been performed. In this test, the value of H is "0″, which represents that it is not statistically significant whereas the value of H is "1"; this represents that it is statistically significant.

8. Open research

The work can be extended to find the sentiments based on other parameters also like **"has attachment"** can be one of the parameters to be included while identifying the sentiment sequence of email threads. In email mining, most of the work has been done in the area of spam filtering, but only a few results have been explored in the field of sentiment analysis of emails. So the researchers will exhibit this area by incorporating the study conducted in this paper as a preliminary step towards analyzing the sentiments of email threads.

Funding information

This research did not receive any specific grant from funding agency in the public, commercial, or not for public sector.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank National Institute of Technology Raipur, Chhattisgarh, India for providing infrastructure and facilities to carry out this research work.

References

Ali, F., Kwak, D., Khan, P., El-Sappagh, S., Ali, A., Ullah, S., … Kwak, K. S. (2019). Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowledge-Based Systems, 174*, 27–42. https://doi.org/10.1016/j.knosys.2019.02.033

Alkhereyf, S., & Rambow, O. (2020). Email classification incorporating social networks and thread structure. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 1336–1345).

AlMahmoud, R. H., Hammo, B., & Faris, H. (2020). A modified bond energy algorithm with fuzzy merging and its application to Arabic text document clustering. *Expert Systems with Applications, 159*, Article 113598. https://doi.org/10.1016/j.eswa.2020.113598

Asani, E., Vahdat-Nejad, H., & Sadri, J. (2021). Restaurant recommender system based on sentiment analysis. *Machine Learning with Applications, 6*, Article 100114.

Baccianella, S., Esuli, A., &Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec* (Vol. 10, No. 2010, pp. 2200-2204).

Balali, A., Faili, H., & Asadpour, M. (2014). A supervised approach to predict the hierarchical structure of conversation threads for comments. *The Scientific World Journal, 2014*. https://doi.org/10.1155/2014/479746

Balali, A., Faili, H., Asadpour, M., & Dehghani, M. (2013). A supervised approach for reconstructing thread structure in comments on blogs and online news agencies. *Computación y Sistemas, 17*(2), 207–217.

Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2021). ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Generation Computer Systems, 115*, 279–294.

Bespalov, D., Qi, Y., Bai, B., & Shokoufandeh, A. (2012). Sentiment classification with supervised sequence embedding. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 159–174). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-33460-3_16.

Bogawar, P. S., & Bhoyar, K. K. (2012). Email mining: A review. *IJCSI International Journal of Computer Science Issues, 9*(1), 429–434.

Carrillo-de-Albornoz, J., & Plaza, L. (2013). An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification. *Journal of the American Society for Information Science and Technology, 64*(8), 1618–1633. https://doi.org/10.1002/asi.22859

Cselle, G., Albrecht, K., & Wattenhofer, R. (2007). BuzzTrack: Topic detection and tracking in email. In *Proceedings of the 12th international conference on Intelligent user interfaces* (pp. 190–197).

Cui, H., Mittal, V., &Datar, M. (2006). Comparative experiments on sentiment classification for online product reviews. In *AAAI* (Vol. 6, No. 1265-1270, p. 30).

Daudert, T. (2021). Exploiting textual and relationship information for fine-grained financial sentiment analysis. *Knowledge-Based Systems, 230*, Article 107389.

Dehghani, M., Asadpour, M., & Shakery, A. (2012). An evolutionary-based method for reconstructing conversation threads in email corpora. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 1132–1137). IEEE. https://doi.org/10.1109/ASONAM.2012.195.

Dehghani, M., Shakery, A., Asadpour, M., & Koushkestani, A. (2013). A learning approach for email conversation thread reconstruction. *Journal of Information Science, 39*(6), 846–863. https://doi.org/10.1177/0165551513494638

El-Din, D. M. (2016). Enhancement bag-of-words model for solving the challenges of sentiment analysis. *International Journal of Advanced Computer Science and Applications, 7*(1).

Ezpeleta, E., Garitano, I., Zurutuza, U., & Hidalgo, J. M. G. (2017). Short Messages Spam Filtering Combining Personality Recognition and Sentiment Analysis. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 25*(Suppl. 2), 175–189. https://doi.org/10.1142/S0218488517400177

Ezpeleta, E., Velez de Mendizabal, I., Hidalgo, J. M. G., & Zurutuza, U. (2020). Novel email spam detection method using sentiment analysis and personality recognition. *Logic Journal of the IGPL, 28*(1), 83–94. https://doi.org/10.1093/jigpal/jzz073

Feng, S., Wang, D., Yu, G., Gao, W., & Wong, K. F. (2011). Extracting common emotions from blogs based on fine-grained sentiment clustering. *Knowledge and Information Systems, 27*(2), 281–302. https://doi.org/10.1007/s10115-010-0325-9

He, Y., & Zhou, D. (2011). Self-training from labeled features for sentiment analysis. *Information Processing & Management, 47*(4), 606–616. https://doi.org/10.1016/j.ipm.2010.11.003

Heerschop, B., Goossen, F., Hogenboom, A., Frasincar, F., Kaymak, U., & de Jong, F. (2011). October). Polarity analysis of texts using discourse structure. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1061–1070). https://doi.org/10.1145/2063576.2063730

Hofmann, T. (2013). Probabilistic latent semantic analysis. *arXiv preprint arXiv: 1301.6705*.

Joshi, S., Contractor, D., Ng, K., Deshpande, P. M., & Hampp, T. (2011). Auto-grouping emails for faster e-discovery. *Proceedings of the VLDB Endowment, 4*(12), 1284–1294. https://doi.org/10.14778/3402755.3402762

Joty, S., Carenini, G., & Lin, C. Y. (2011). Unsupervised modeling of dialog acts in asynchronous conversations. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence* (Vol. 22, No. 3, p. 1807).

Kooti, F., Lerman, K., Aiello, L. M., Grbovic, M., Djuric, N., & Radosavljevic, V. (2016). Portrait of an online shopper: Understanding and predicting consumer behavior. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (pp. 205–214). https://doi.org/10.1145/2835776.2835831

Liu, S., & Lee, I. (2018). Discovering sentiment sequence within email data through trajectory representation. *Expert Systems with Applications, 99*, 1–11. https://doi.org/10.1016/j.eswa.2018.01.026

Liu, S., Lee, K., & Lee, I. (2020). Document-level multi-topic sentiment classification of Email data with BiLSTM and data augmentation. *Knowledge-Based Systems, 105918*. https://doi.org/10.1016/j.knosys.2020.105918

Ma, S., Sun, X., Wang, Y., & Lin, J. (2018). Bag-of-words as target for neural machine translation. *arXiv preprint arXiv:1805.04871*. doi:10.18653/v1/P18-2053.

Mao, Y., & Lebanon, G. (2007). Isotonic conditional random fields and local sentiment flow. In *Advances in neural information processing systems* (pp. 961-968).

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mesnil, G., Mikolov, T., Ranzato, M. A., &Bengio, Y. (2014). Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. *arXiv preprint arXiv:1412.5335*.

Nagwani, N. K., & Sharaff, A. (2017). SMS spam filtering and thread identification using bi-level text classification and clustering techniques. *Journal of Information Science, 43*(1), 75–87. https://doi.org/10.1177/0165551515616310

Nascimento, M. C., & De Carvalho, A. C. (2011). Spectral methods for graph clustering–a survey. *European Journal of Operational Research, 211*(2), 221–231. https://doi.org/10.1016/j.ejor.2010.08.012

Naseem, U., Razzak, I., Musial, K., & Imran, M. (2020). Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems, 113*, 58–69. https://doi.org/10.1016/j.future.2020.06.050

Nenkova, A., &Bagga, A. (2004). Facilitating email thread access by extractive summary generation.*Recent advances in natural language processing III: selected papers from RANLP, 2003*, 287-294.

Niu, L., & Shi, Y. (2010). Semi-supervised plsa for document clustering. In *2010 IEEE International Conference on Data Mining Workshops* (pp. 1196–1203).

Pang, B., Lee, L., &Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

Qiu, G., He, X., Zhang, F., Shi, Y., Bu, J., & Chen, C. (2010). DASA: Dissatisfaction-oriented advertising based on sentiment analysis. *Expert Systems with Applications, 37*(9), 6182–6191. https://doi.org/10.1016/j.eswa.2010.02.109

Ren, Y., Wang, R., & Ji, D. (2016). A topic-enhanced word embedding for Twitter sentiment classification. *Information Sciences, 369*, 188–198. https://doi.org/10.1016/j.ins.2016.06.040

Rezaeinia, S. M., Rahmani, R., Ghodsi, A., & Veisi, H. (2019). Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications, 117*, 139–147. https://doi.org/10.1016/j.eswa.2018.08.044

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management, 52*(1), 5–19. https://doi.org/10.1016/j.ipm.2015.01.005

Severyn, A., & Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 959–962). https://doi.org/10.1145/2766462.2767830

Sharaff, A., & Nagwani, N. K. (2016). Email thread identification using latent Dirichlet allocation and non-negative matrix factorization based clustering techniques. *Journal of Information Science, 42*(2), 200–212. https://doi.org/10.1177/0165551515587854

Sharaff, A., & Nagwani, N. K. (2020). ML-EC2: An Algorithm for Multi-Label Email Classification Using Clustering. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 15*(2), 19–33. https://doi.org/10.4018/IJWLTT.2020040102

Sharaff, A., & Soni, A. (2018). Analyzing Sentiments of Product Reviews Based on Features. In *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 710–713).

Shen, J., Brdiczka, O., & Liu, J. (2013). Understanding email writers: Personality prediction from email messages. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 318–330). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-38844-6_29.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., & Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 151–161).

Srinivasarao, U., & Sharaff, A. (2021a). Email Sentiment Classification Using Lexicon-Based Opinion Labeling. In *Intelligent Computing and Communication Systems* (pp. 211–218). Singapore: Springer.

Srinivasarao, U., & Sharaff, A. (2021b). Sentiment analysis from email pattern using feature selection algorithm. *Expert Systems, e12867*. https://doi.org/10.1111/exsy.12867

Tai, C. H., Tan, Z. H., Lin, Y. S., & Chang, Y. S. (2015). Mental disorder detection and measurement using latent Dirichlet allocation and SentiWordNet. In *2015 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 1215–1220). IEEE. https://doi.org/10.1109/SMC.2015.217.

Tang, D., Qin, B., & Liu, T. (2015). Learning semantic representations of users and products for document level sentiment classification. 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. 1014–1023). https://doi.org/10.3115/v1/P15-1098

Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems, 24*(7), 1024–1032. https://doi.org/10.1016/j.knosys.2011.04.014

Valdivia, A., Luzón, M. V., Cambria, E., & Herrera, F. (2018). Consensus vote models for detecting and filtering neutrality in sentiment analysis. *Information Fusion, 44*, 126–135. https://doi.org/10.1016/j.inffus.2018.03.007

Vashishtha, S., & Susan, S. (2019). Fuzzy rule based unsupervised sentiment analysis from social media posts. *Expert Systems with Applications, 138*, Article 112834. https://doi.org/10.1016/j.eswa.2019.112834

Wang, G., Zhang, Z., Sun, J., Yang, S., & Larson, C. A. (2015). POS-RS: A Random Subspace method for sentiment classification based on part-of-speech analysis. *Information Processing & Management, 51*(4), 458–479. https://doi.org/10.1016/j.ipm.2014.09.004

Wei, C. P., & Chang, Y. H. (2007). Discovering event evolution patterns from document sequences. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 37*(2), 273–283. https://doi.org/10.1109/TSMCA.2006.886377

Wen, Q., Gloor, P. A., FronzettiColladon, A., Tickoo, P., & Joshi, T. (2020). Finding top performers through email patterns analysis. *Journal of Information Science, 46*(4), 508–527. https://doi.org/10.1177/0165551519849519

Wu, Y., & Oard, D. W. (2005). Indexing emails and email threads for retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 665–666).

Xie, X., Ge, S., Hu, F., Xie, M., & Jiang, N. (2019). An improved algorithm for sentiment analysis based on maximum entropy. *Soft Computing, 23*(2), 599–611. https://doi.org/10.1007/s00500-017-2904-0

Xie, H., Lin, W., Lin, S., Wang, J., & Yu, L. C. (2021). A multi-dimensional relation model for dimensional sentiment analysis. *Information Sciences*.

Yousefpour, A., Ibrahim, R., Hamed, H. N. A., & Yokoi, T. (2016). Integrated feature selection methods using metaheuristic algorithms for sentiment analysis. In *Asian Conference on Intelligent Information and Database Systems* (pp. 129–140). Berlin, Heidelberg: Springer.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011, 89*.

Zhao, J., Liu, K., & Xu, L. (2016). Sentiment analysis: mining opinions, sentiments, and emotions.

Zhou, L., Ye, W., Plant, C., & Böhm, C. (2017). Knowledge discovery of complex data using Gaussian mixture models. In *International Conference on Big Data Analytics and Knowledge Discovery* (pp. 409–423). Cham: Springer. https://doi.org/10.1007/978-3-319-64283-3_30.