Knowledge-Based Systems 208 (2020) 106443

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email

Zahid Halim *, Mehwish Waqar, Madiha Tahir

The Machine Intelligence Research Group (MInG), Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, 23460, Pakistan

ARTICLE INFO

Article history:
Received 21 March 2020
Received in revised form 19 August 2020
Accepted 1 September 2020
Available online 19 September 2020

Keywords:
Emotion recognition
Email analysis
Supervised learning
Sentiment analysis

ABSTRACT

Identification of emotion hidden in limited text is an active research problem. This work presents a framework for the same using email text. The present work is based on machine learning methods and utilizes three classifiers and three feature selection methods. The novelty of the proposed framework is the utilization of in-text features to identify emotion contained in short texts and development of a dataset for this purpose. Six emotions, namely, *neutral*, *happy*, *sad*, *angry*, *positively surprised*, and *negatively surprised* are utilized here based on baseline theories on human emotion. Experiments are performed on three datasets including a benchmark and one local dataset. These experiments are performed by extracting 14 in-text features from the data. The proposed framework is evaluated using four standard evaluation metrics. Based on the feature selection results, experiments are performed on the datasets under consideration by vertically partitioning them into all features, top features, and bottom features. Qualitative and quantitative comparison of the proposed work is also made with two state-of-the-art methods. The obtained results suggest better performance of the current work with an average accuracy of 83%. The proposed framework can be utilized in an assortment of domains to identify human emotion by providing limited text as an input.

1. Introduction

Identification of human emotion for effective human computer interaction and informed decision-making is an active area of Artificial Intelligence (AI) research. However, it is a challenging task for the AI community to identify affective state(s) of a person using her text samples only. Humans generate their digital footprint in multiple ways through diverse computing devices. This is because the advances in information, communication, and technology has transformed the current era into a digital world. Computers and smartphones have become an integral part of one's daily life and many routine chores revolve around them [1,2]. This asks for introduction of novel methods to make Human Computer Interaction (HCI) emotionally intelligent [3]. At present, the computing devices are intelligent in different ways, for example, they can recognize speech, can identify the force of a keystroke, can predict the next word while typing text, and can identify body posture. However, as yet, the today's computing devices cannot understand the human emotion associated to or hidden within each kind of input they receive. Empowering computing devices to perceive the human feeling make them fit for downplaying the semantics of the correspondence that occurs among humans and them. This will empower the present data processing gadgets to react fittingly or perhaps more humanly to the instructions they receive. Affective computing therefore is the application of pattern recognition that deals with enabling computers to be emotionally intelligent. Emotion recognition can be done by using data obtained from facial expressions, hand gestures, body language, blood pressure, heartbeat rate, body temperate, or voice tone, to name a few. However, to identify hidden emotions in text, Machine Learning (ML) methods can play an important role [4]. A couple of past works demonstrate proof of concept studies regarding this [5].

Identification of human emotions from written text and user-generated contents can be utilized for the improvement of the client cooperation in an assortment of settings. Research on the discovery of feelings from messages, tweets, short text, and video remarks (on YouTube and so forth) is utilized to enhance the clients' experience for quality improvement. Because of the ever growing usage of the Internet, sentiment analysis using text has become an emerging area of reach these days [6]. The domain is more commonly referred to as opinion mining. It is generally used for business intelligence to analyze the behavior of public and users towards a company's product and various brands. On various e-commerce websites, different options are available to obtain feedback, comments, and reviews from the

* Corresponding author.
E-mail addresses: zahid.halim@giki.edu.pk (Z. Halim), gcs1635@giki.edu.pk (M. Waqar), madiha.tahir@giki.edu.pk (M. Tahir).

https://doi.org/10.1016/j.knosys.2020.106443
0950-7051/

users. Users provide feedback according to their intellectual level and sentiments. This feedback is recorded, analyzed and is used to improve the product/service, organizational decisions and also organizational services [7]. Sentiment analysis of the text is also used by the intelligence agencies to know about the prospective security threats. Intelligence agencies analyze general public views regarding infrastructure, administration, institutions, and services [8]. The political behavior of the people can be extracted through their opinions on different social networking websites for a specific candidate or a particular political party. Other applications include product marketing, social behavior judgment, political perception analysis, and brand performance. Social networking websites provides an opportunity to give positive or negative reviews about a topic. People feel more comfortable and free to express their views on social networking sites and discuss various issues within/outside their social circle.

Generally, sentiment analysis and the emotion recognition are regarded as the same. However, they are fundamentally dissimilar. The process of analyzing the attitude in a given text, speech or writing with respect to a specific topic is sentiment analysis [9]. The same is referred to as opinion mining while dealing with product reviews. For sentiment analysis, the basic task is to categorize the text into three classes, namely, *positive*, *negative* or *neutral*. Whereas, extracting and identifying a specific human emotion (e.g., *happy*, *sad*, *surprised*, or *angry*) is the key task to be performed by an emotion recognition system. Emotion recognition, therefore, is a subdomain of sentiment analysis at a lower level of abstraction with an aim to get a better understanding of the human emotion. Liu et al. [10] has previously addressed the issue of emotion identification. However, identification of a dominant emotion in limited text, like emails, is a challenging task requiring novel methods to address this issue. Many theories exist on the basic human emotions. The widely accepted of these is the Paul Ekman's theory according to which there are six basic human emotions, namely, *happy*, *sad*, *fear*, *disgusting*, *angry*, and *surprise* [11]. The same is adopted here for deciding the number and type of emotions to be considered in this study.

1.1. Open issues

One of the challenges in emotion identification from the email text is the availability of datasets with limited size instead of large data or multiple documents which do not reflect the actual email contents. Therefore, emotion recognition from short text, like that of an email has many open issues. The traditional document classification methods, like, bag-of-words analysis fail here due to fewer number of words/sentences [12]. It provides sparsity in features due to which machine learning algorithms and text mining methods do not perform well [13]. Sparsity and span over words make it challenging for the classification techniques and clustering algorithms to give better results. Many approaches have been introduced in the past to classify short text as *neutral*, *positive* or *negative* [14,15]. Few tools are also available to process text and identify the polarity of text given three options, i.e., *positive*, *negative* or *neutral* [16]. However, all of this majorly belong to the domain of sentiment analysis.

Email is a specific form of short text which is used as personal or organizational communication channel across the world. It is used as professional mean of communication due to its high speed, popularity, and trustful security. Emotion recognition from an email's text can be helpful on many levels. Recognition of the dominant emotion contained in an email can directly influence one's decisions related to the organization or on a personal level. To classify emails as *neutral*, *positive* or *negative*, a few available tools and approaches can be used [16]. However, extracting a specific emotion out of an email, for example, *angry*, *surprised*, *happy* or *sad* still remain an open challenge for the AI community.

1.2. Scope and applications

The scope of this work is to identify dominant emotion in an email using its in-text features only. An email can consist of text, images, web links, videos, and other attachments. However, this work focuses only on the textual data of an email for feature extraction and emotion identification. For this, 14 features are extracted from the email. Later, three feature selection techniques are applied to these attributes, namely, Principal Component Analysis (PCA), Mutual Information (MI), and Information Gain (IG). The resulting features are then used by the classification algorithms for learning. The proposed approach can be used as add-on with a web browser, as a stand-alone software, or as a satellite feature with the email services to extract a specific emotion using its text. The current proposal aims at utilize the supervised learning methods for the problem at hand. The email is a mode of communication for official correspondence and for personal interaction. Identification of the dominant emotion in a particular email will help in appropriate interpretation at the receiving end. This will also have other benefits from a business' point of view, like, targeted advertisement, emotional therapy, and many others. Additionally, emotion recognition can be used by email platforms for spam filtering to separate emails which may contain offensive data of any form. The present work considers the emotions identified in the Paul Ekman's theory. He identified six basic emotions according to isolated culture of people from the Fori tribe in Papua New Guinea in the year 1972 [11]. Therefore, the present work has adopted the same number. Additionally, the majority of the past works like, [5,17–19] also use six number of emotions in their studies. A few past contributions, like [20] has used eight emotions, namely, *happy*, *calm*, *in love*, *positive surprise*, *negative surprise*, *angry*, *sad*, and *afraid*. The two additional emotions considered in [20] are *calm* and *in love*. Since these two are not mentioned in the Paul Ekman's theory as basic emotions and inducing these two is also challenging, therefore the present work continues with the six emotions based on the majority of the past works and utilizes the standard theory on human emotion.

1.3. Our contributions and novelty

This work presents a supervised learning solution for emotion recognition in email texts. The proposed approach categorizes an email into one of the six emotions, i.e., *neutral*, *happy*, *sad*, *angry*, *positively surprised*, and *negatively surprised*. This is done based on the analysis performed on the email's text only. For experiments in this work, three datasets are used. First dataset is the Enron email database [21] which contain emails of about 150 senior management officers. This email corpus contains approximately 0.5 million emails and is publicly available online for research purposes. However, the dataset is not labeled, therefore, the labeling task is carried out for this dataset in this work. Identifying emotions from small text is a challenging task and to the best of our knowledge a dataset that contains labeled emails with all abovementioned emotions is not publicly available. Due to this limitation, the current work creates such dataset for the utilization by the pattern learning module. For the second dataset, 60 participants are engaged and stimuli are induced to create an environment specific to each of the six emotions. Later, they are instructed to write emails on a specific topic while they are under the influence of the induced emotion. This dataset contains 339 emails (after discarding a few samples due to poor quality). The third dataset is the combined form of the abovementioned two datasets. There are generally three types of approaches to extract emotions from short text, i.e., machine learning algorithms, hybrid method, and lexicon method [22].

Machine learning algorithms use distinct features to classify or cluster text. The hybrid method combines machine learning algorithms and lexicon methods. Lexicon method uses the lexicon of sentiment, i.e., it uses a collection of specific terms available in the sentiments. These terms can be of two types, i.e., corpus-based and dictionary-based [23]. This work utilizes supervised learning for emotion recognition based on email's text and the unsupervised learning methods are used here for verifying the results. For supervised learning, three classifiers are used, namely, Artificial Neural Network (ANN), Random Forest (RF), and Support Vector Machines (SVM). Before providing the classifiers with the datasets for training and testing, the dataset is vertically partitioned into three subsets; all features, top features, and bottom features. For this, three feature selection methods are utilized, namely, Principal Component Analysis (PCA), Mutual Information (MI), and Information Gain (IG). For evaluating the performance of the classifiers, four evaluation metrics are used, i.e., accuracy, F1 score, precision, and recall. To validate the classification results, unsupervised learning technique (i.e., clustering) is utilized. To identify the number of disjoint groups captured in the two datasets, three clustering methods are used in this work, namely, *k*-means, *k*-medoids, and fuzzy c-means. The clustering results are evaluated using four cluster validity indices, namely, Davies Bouldin Index (DBI), Dunn Index (DI), Silhouette Coefficient (SC), and Gap Evaluation (GE). Based on the abovementioned details, the key contributions of this work are as follows.

- Presentation of a machine learning-based framework to identify dominant emotion hidden in email text.
- Utilization of in-text features only for the prediction of emotion.
- Contribution of a novel balanced and labeled dataset containing six basic human emotions using email text to be utilized by the AI community to train existing and build new models for the emotionally intelligent computing systems.
- Another novelty of the dataset contributed by this work is its non-posed nature.
- Evaluation of the local dataset using the clustering methods for verifying the presence of six human emotions.
- Assessment of the proposed framework using four standard evaluation metrics.
- A method enabling the email service providers to have an add-on for identifying the dominant emotion hidden in an email for appropriate interpretation.

The key novelty of this work is the utilization of in-text features to identify the dominant emotion in the email text. There are a few past contributions based on deep learning methods, like [17] and [5], which take the complete sample as one batch and predict the emotion. The limitation of such methods is that they do not let the user know specifically what data has been used for the prediction purpose and which attributes play vital role in the prediction task. Such methods are a black box and there is nothing much for the user to manipulate. Whereas, the present work provides the user with the flexibility to select and extract a variety of features from the dataset and experiment with them to see which ones perform better. Another innovative aspect of the present work is the contribution of a labeled dataset consisting of short text in the form of emails for the research community. In addition to the utility of a benchmark dataset, a local data is also captured for experiments. An experiment is performed here to create another labeled dataset. Sixty volunteers were engaged for this purpose and a stimulus was induced into the participants for each of the six emotions under consideration in this work. Later, the participants were asked to write an email on a specific subject for each of the emotion types. This provided us with 339 emails. An advantage of this dataset is it being balanced by having the same number of samples of all classes of emotions. The collected dataset is also verified using the unsupervised learning methods, i.e., clustering to make sure that the six categories of emotions under consideration in this work are actually available in the collected data. The assessment of various features of the proposed framework using multiple techniques is also an innovative feature. For example, the past works usually utilize only one feature selection technique for the extraction of attributes from the data. However, it is a known fact that the features extracted using one feature selection method may differ than the ones extracted using a different feature selection technique. Therefore, this work incorporates in the proposed framework three feature selection techniques and the final feature set is opted based on the majority vote.

The rest of the paper is organized as follows. Section 2 lists the previous work on emotion recognition from short text, i.e., Twitter data, Facebook data, and emails. The section covers various methods based on machine learning and lexicon-based approaches. Section 3 list details about the datasets. Section 4 presents the proposed approach by explaining participants' details, data recording methodology, preprocessing, feature selection, classifiers, clustering, and validation techniques. Section 5 present experiments and their results. Section 6 contains the discussion and finally, Section 7 concludes this work.

2. Related work

Sentiment analysis is an active domain of research in computer science. This can further be categorized into the identification of emotions at a lower level of abstraction. Generally, sentiments can either be *positive* or *negative*. However, the emotions can have six categories, i.e., *sad*, *happy*, *neutral*, *positively surprised*, *negatively surprises*, and *angry* [20]. This section covers the literature related to the sentiment analysis and emotion recognition.

2.1. Sentiment analysis using short text

Twitter is one of the popular social networks where users share views about various events/things and express their sentiments. The authors in [24] show the modeling of Twitter data for prediction purpose using topic-based context and social context. Previously, work has been done to extract sentiment of opinions from the users' tweets and then to use it for various analysis. However, utilizing opinion and contextual information to predict the users' opinion remain a challenging task. The work in [24] presents a framework named Social context and Topical context incorporated Matrix Factorization (ScTCMF) to measure the content-based correlation between topics and applied the Topic Content Similarity (TCS). During experiments, they evaluate ScTCMF framework using real-world data and show that both types of contexts, i.e., topic-based context and social context are beneficial for extracting sentiments from the users' opinion text.

Machine learning still uses Bag of Words (BoW) approach for text analysis to predict sentiments. The BoW approach has many limitations while performing sentiment analysis. The work in [25] presents a model called Dual Sentiment Analysis (DSA). This model is used for sentiment classification. Firstly, reversed data expansion is performed for testing and training purpose. Later, the DSA uses original and reversed reviews to classify them into sentiments. In the first classification approach, two sentiment classes are defined, i.e., *positive* and *negative*. In the second classification task, three sentiments are used as class labels, i.e., *positive*, *negative*, and *neutral*. The work uses corpus method pseudo-antonym dictionary. The authors in [26] focuses on students' Twitter posts for sentiment analysis and use it to

Table 1
Summary of the past work on emotion recognition using short text.

Works	Data domain	No. of emotions	Method	Dataset
Kratzwald et al. [17]	Social network	4	RANN with transfer learning	Multiple social network dataset
Tripathi et al. [5]	Text/speech/motion	4	Machine learning, Artificial neural network	IEMOCAP
Matsumoto et al. [18]	Text	8	Bag of Concepts using *k*-Nearest Neighbors and Maximum Entropy Method	YSEC
Amelia et al. [19]	Short stories	6	Hybrid method (SVM, naïve Bayes testimonial, Multinomial Logistic Regression)	Short stories dataset
Ren et al. [24]	Twitter	3	ScTcMF Framework	Real-world Twitter data using Twitter API
Xia et al. [25]	Twitter	3	Dual sentiment analysis	Nine datasets on product reviews
Halim et al. (proposed)	Email text	6	Machine learning-based framework	Enron email database and local data

get students' learning experiences about education. Their proposal is based on the integration of two techniques, i.e., data mining and qualitative analysis. Engineering students' Twitter posts are collected to get their sentiments regarding studies, the learning process in the educational department, lectures, and a few other aspects. This results in the creation of two datasets having approximately 25k–35k tweets. Using this data, their work identifies various issues faced by the engineering students. Multi-classification method of machine learning is used for categorization of the tweets. Precisely, six classes are identified in [26] based on students' sentiments, namely, heavy study load, lack of social engagement, negative emotion, sleep problems, diversity issues, and "others".

The work in [27] presents an approach for cross-domain data which uses a Sentiment Sensitive Thesaurus (SST). In their approach, on one hand, the distributional sentiment sensitive thesaurus is produced using labeled and unlabeled data. For source domains, SST of the labeled data is produced. On the other hand, for the target and source domains, the SST of unlabeled data is formed. The utilization of SST produces better results and show almost similar sentiments as identified by SentiwordNet,[1] a lexical resource for opinion mining. The effect of different types of adverbs on sentiment analysis is studied in [28]. Hadoop is used with classification. The work considers eight types of adverbs. An amazon-based dataset having around 60k reviews is used which is a labeled dataset. Results show that for positive and negative sentiments of the opinions, locative adverbs (RL) and preposition adverbs (RP) have higher impact on results. Whereas, the general comparative adverb (RRR) has an impact on neutral sentiment classification.

2.2. Emotion recognition using short text

A method is introduced in [17] to use the Recurrent Artificial Neural Network (RANN) and transfer learning for emotion recognition. The results verify that transfer learning and RANN both perform better than the general methods of machine learning. Another neural network enhancement for emotion recognition is proposed in [5]. The work uses a neural network as a classifier to predict multi-model emotion recognition using IEMOCAP dataset.[2] The dataset is constructed using motion, speech, and text. For motion data, hand movement, facial expressions, and rotation are considered. To extract emotion from three different types of data, separate approaches are introduced. Text-based emotion recognition uses three models. The first model uses 32 filters with Rectified Linear Unit (ReLU) activation function having 256 dimensions. The second one uses Long Short-Term Memory (LSTM) multiple layers and activation function ReLU having 256 and 512 dimensions, respectively. While the third model is constructed using the same specifications as the second model, however, it is assembled using 128 dimensions. The three models have approximately 62%, 64.68%, and 64.78% accuracy, respectively.

Emotion recognition is mostly done using dictionary-based words. The author in [18] introduces an approach called the Bag-of-Concepts which considers new words on the web and uses them for emotion recognition. They create a dataset using weblog text articles. Another Twitter dataset is used in [18] to evaluate sensibility of words in the context of Twitter posts. Multiple classifiers including *k*-Nearest Neighbors (*k*-NN) and Maximum Entropy Method (MEM) are applied to identify one of the seven emotions, namely, *anxiety*, *love*, *sorrow*, *surprise*, *pleasure*, *hate*, and *hope*. Short stories can have many emotions hidden in them. Extraction of only one dominant emotion out of the six possibilities is done in [19]. The six emotions are selected on the basis of Paul Ekman's proposed emotion categorization [11] which includes *angry*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*. A learning-based model comprising of three classifiers, i.e., naïve Bayes multinomial, Support Vector Machine (SVM), and Multinomial Logistic Regression (MLR) is used. This hybrid approach recognizes one dominant emotion and gives approximately 63.33% average accuracy. Table 1 summarizes the past works on emotion recognition and sentiment analysis using short text.

The work in [29] presents an approach for the classification of emotion in the context of Open Source Software (OSS). They collect the data from Stack Overflow posts and a dataset obtained from JIRA issue tracker comments. For the prediction task, HOMER and RAkEL models are used in their work. Their method achieves an F1 score up to 81.1%. Their work utilizes six classes, namely, *joy*, *love*, *anger*, *surprise*, *sadness*, and *fear*. The work in [30] presents an application of Emotional Text Mining (ETM) in the field of brand management. Their approach is primarily used for the profiling of social media users. The data for experiments is extracted from Twitter regarding an unknown sports brand. The work in [31] proposes to utilize cell phones as the gadget that brings together relevant data of the older individuals, concentrating on feeling acknowledgment. Their proposal includes the utilization of image processing [32,33] techniques. The proposal in [34] presents a graph-based technique [35] named graph convolutional broad network for the identification of emotion. Their approach use convolutional layer to retrieve graph features and stacks different customary convolutional layers to separate the most relevant attributes [36]. From the application point of view,

1 http://sentiwordnet.isti.cnr.it/
2 https://sail.usc.edu/iemocap/

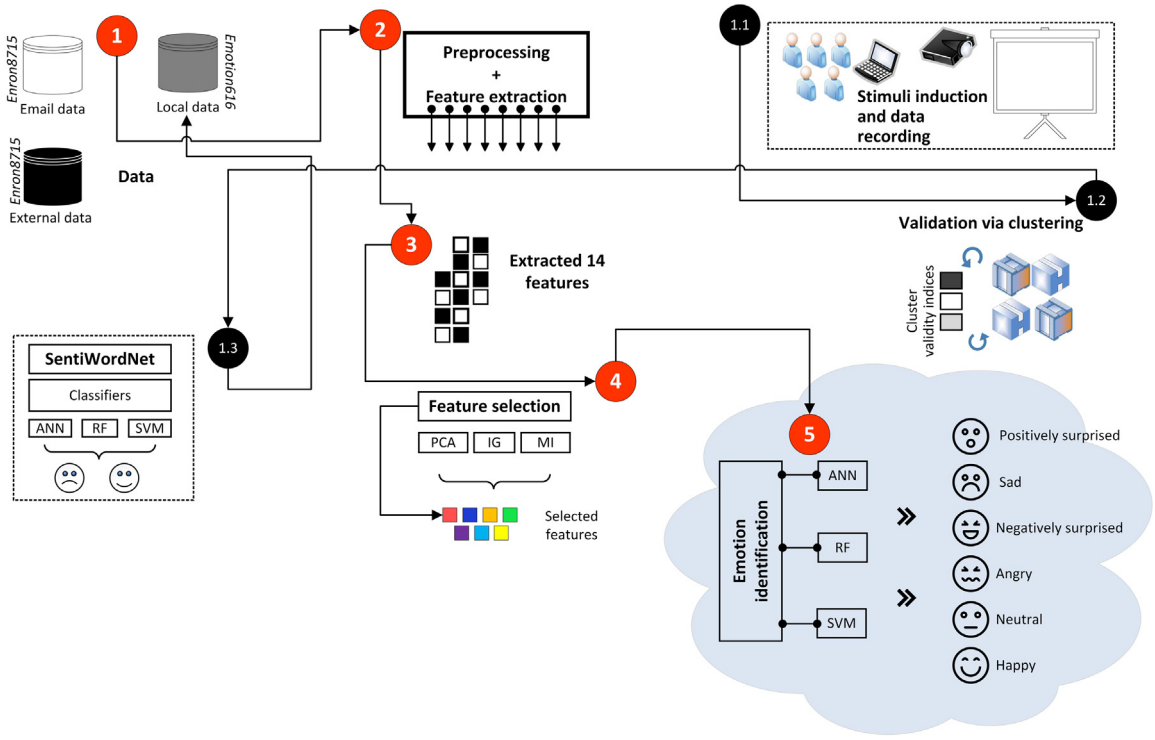

Fig. 1. Overview of the proposed system.

the work in [37] present an innovative application of the emotion recognition methods in borderline personality disorder [38].

2.3. Sentiment analysis using short text of the emails

Sentiment analysis of short email text is studied in [39]. Their work presents a hybrid framework with two algorithms for sentiment analysis. These two algorithms are *k*-means and Support Vector Machines (SVM). Results are compared using *k*-means-based labeling, polarity-based labeling, SentiWordNet-based labeling, SVM, Decision Tree (DT), Naïve Bayes (NB), OneR, and logistic regression. Results show better accuracy of the framework in contrast to other competitors. In the context of identifying dominant emotion hidden in an email, the work in [39] is the closely related past contribution based on sentiment classification. It extracts polarity from email and classifies them in one of the three categories, i.e., *positive*, *negative* or *neutral*. However, to extract a specific emotion from email, it requires more processing and a different method. The current proposal addresses this limitation of the past work.

3. Datasets

As mentioned earlier (and to the best of our knowledge), there are no such publicly available labeled email datasets that could be utilized here. Therefore, an important aim of the present work and also a key contribution is to develop a labeled dataset that contains email text data related to all basic human emotions. This work contributes a novel labeled dataset extracted from email text containing of six basic human emotions. It will be available to the AI community to train existing and build new models for the emotionally intelligent computing systems. The novelty of the dataset here is that it consists of emotion related short text data in the form of email text. For the sake of completeness, the present proposal is also evaluated on an existing benchmark (regarding email data) which is manually labeled using majority vote of three annotators. The local dataset is labeled as *Emotion616* which is collected from 61 volunteers.

Three email datasets are utilized in this work. The first dataset is the Enron email dataset which is a publicly available benchmark (unlabeled) [40]. In this dataset, the emails are collected and organized into folders from 150 people of Enron's senior management. This dataset has around 500k emails. Out of 150 user's folder, first 87 folders were selected in this work where, on an average, 15 emails are selected per folder. These 15 emails are selected from the sent items of each user's folder to create diversity in the emails. Multiple users within an organization can have similar emails in their inbox, however, the sent items generally differ. The labeling of these emails is done manually utilizing majority vote of three annotators. Each email is assigned one of the six labels, i.e., *neutral*, *happy*, *sad*, *angry*, *positively surprised*, and *negatively surprised*. Eventually, 1000 emails out of 1305 are finalized based on the labeling quality. This dataset is named as *Enron8715* in the current proposal comprising of three parts. "Enron" is the first part which means this dataset is a subpart of the Enron email dataset. The digits "87" represents the number of the user's folders selected out of 150 folders for email selection. The digits "15" indicates that 15 emails are selected from each folder. Distribution of the human emotion labels in email data of *Enron8715* is shown in Fig. 2. Number of instances of the *neutral* and *happy* labels are more in number as compared to others. Due to this imbalance between various class labels, another dataset is created here.

The second dataset is entitled as *Emotion616* which is collected by inducing stimuli into 61 volunteers. The title *Emotion616* has three parts. The first part "Emotion" indicates that this dataset has output class labels based on human emotions. The second part "61" indicates the number of participants engaged. The last fragment, i.e., "6" show the number of class labels. After discarding a few emails based on quality, the dataset *Emotion616* has 339 emails remaining, six emails per participant (one email corresponding to each of the emotion). Fig. 2 shows the emotion label distribution in *Emotion616* dataset. Almost all emotion classes are

Table 2
Videos used for stimulus induction.

Video	URL	Emotion label	Views (as of May 2020)
A Foreigner's Trip from London to Gilgit Baltistan Pakistan	https://www.youtube.com/watch?v=nXwijt1EUDo	Happy	0.47 million
The Bermuda Triangle Mystery Has Been Solved	https://www.youtube.com/watch?v=q_5n7URd2Gk	Positively surprised	26.46 million
MAN	https://www.youtube.com/watch?v=WfGMYdalClU	Negatively surprised	41.49 million
Heart Touching - Quit smoking (inspirational video) - Save life	https://www.youtube.com/watch?v=5qeWUCPBVXA&t=12s	Neutral	0.046 million
Heart Touching Children In Syrian Civil War: Share If You Care	https://www.youtube.com/watch?v=jdKHVnHTXkU	Angry+ Sad	1.02 million

Table 3
Datasets.

Dataset	Number of participants	Source	Number of emails	Emotion classes
Enron8715	87	Enron organization	1000	6
Emotion616	61	Survey through emotion induction	339	6
Emotion135106	147	Enron organization + Survey	1351	6

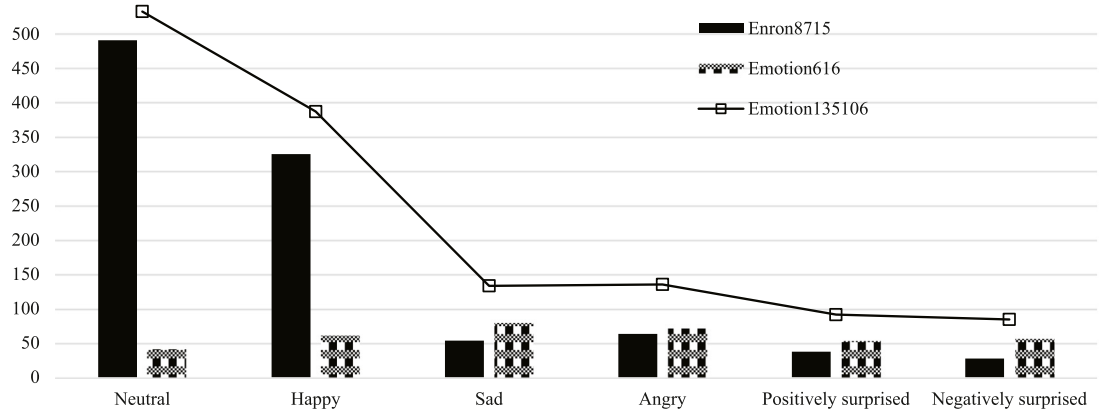

Fig. 2. Distribution of the six emotions in three datasets.

equally distributed in this dataset. Emails having *sad* emotion label are highest in number. For the creation of this dataset, the environment for a specific emotion was created through displaying videos selected from the YouTube based on the comments and views (in millions). Table 2 lists key features of the videos utilized in this work to induce the stimuli. A survey form was distributed among participants to collect demographic information. The data was collected in three sessions. During each session, the videos were played one at a time and at the end of each video, a question statement was provided to the participants instructing them to write an email in a word document in reply to the asked question. The third dataset is labeled as *Emotion135106* which is a combined dataset using *Enron8715* and *Emotion616*. The name *Emotion135106* also consists of three parts. "Emotion" indicates that this dataset contains data which has human emotion-based output class labels. The digits "1351" indicate the number of emails and "06" shows the number of output classes. Fig. 2 also shows the distribution of emotion class labels in the Emotion135106 dataset. Table 3 lists the summary of the three datasets utilized in this work.

4. Proposed solution

The proposed solution for identifying the dominant emotion in the email text is explained in this section. This work presents a machine learning-based solution for the identification of dominant emotion in an email using its in-text data only. For this purpose, a complete framework is devised here. Although the proposed framework theoretically starts with the data input and its preprocessing steps, however, the creation of a balanced and labeled dataset covering the six basic emotions mentioned in the Paul Ekman's theory is also an aim of this work. This is done because, to the best of our knowledge, there are no publicly available datasets having emails labeled with their dominant emotion. For the sake of completeness, the present proposal is also evaluated on an existing benchmark which is manually labeled using majority vote of three annotators. The local dataset is labeled as *Emotion616* which is collected from 61 volunteers. Overall, the proposed framework has four phases, data collection being the first one. In the second phase, the data is preprocessed. This phase is executed irrespective of the fact that the data is being fed into the framework internally or it is coming from external sources. The preprocessing phase performs tokenization, redundancy elimination, and Parts of Speech (PoS) tagging. The third phase is the feature extraction and selection task. Here, 14 predefined in-text features are extracted from each email and three feature selection techniques are used for feature ranking and opting for the optimal feature set. The fourth (and final) phase of the framework is the classification task, where three classifiers are used for predicting the dominant emotion in the email text. In addition to the four phases of this framework, the local data collected in this work also needs to be validated. This is an important step to make sure the availability of the six basic emotions in the local dataset. This will ensure better training of the learning module (i.e., the fourth phase of this framework). For this, a separate module based on two unsupervised learning methods (i.e., clustering) and multiple cluster validity indices is designed. The proposed framework is incorporated with three

feature selection methods, three classification approaches, two clustering techniques, and multiple cluster validity indices. This is done for rigorousness and generality (to the extent possible) of the proposed work. For other datasets, the users may opt for any one technique in each of the phases of this framework. However, utility of multiple techniques in each phase enables to obtain generalized results. Fig. 1 visually shows the overall working of the proposed solution. Details about the four phases of the proposed framework are listed in the following. This section also lists the clustering algorithms utilized in this proposal for data validation.

4.1. Pre-processing

The email data considered in this work is of textual nature only. Therefore, a number of preprocessing steps are required to prepare the data for further utilization by the pattern recognition modules. Four preprocessing steps are used in this work. The first step is the tokenization of the email into separate words. This enables to find frequencies of each of the words and discard any redundant or unwanted portions. The second preprocessing step involves stop words removal from the list of tokens. The third step is lemmatization of the remaining tokens. This process utilizes PoS and context for conversion of the word into a root word. This enables to obtain the base/dictionary form of a word. Finally, the PoS tagging is done for further processing. The PoS assigns each word a category based on its syntactic functions. In this work the tag CC is used for conjunction, JJ, JJR, and JJS are used for adjective, for verbs VB and VBD are utilized, adverbs are tagged by RB, RBS, and RBR, whereas for noun NN, NNP, and NNS are utilized.

4.2. Feature extraction and selection

This work has utilized 14 features to be extracted from the email text to enable the learning module to identify the underlying emotion hidden in the emails. Identification of the 14 features is based on the work in [41]. These features include, number of verbs, number of adjectives, number of adverbs, number of conjunction, number of nouns, five Boolean variables indicating availability of verbs, adjectives, adverbs, conjunctions and nouns in an email, two variables representing counts of the punctuation marks "?" and "!", number of positive words, and the number of negative words. The feature extraction from the email text is done using python and the Natural Language Toolkit (NLTK).[3] Once all the features are extracted from the email text, the next step is to apply feature selection techniques on these so that their effectiveness can be determined. Three feature selection techniques are used in the proposed framework, i.e., Principal Component Analysis (PCA), Information Gain (IG), and Mutual Information (MI). PCA is chosen because it joins comparatively similar components to make new ones, better than the original one. MI generally considers dependencies between features and class, as compared to statistical methods, like PCA, that do not consider class labels. It provides superior feature extraction than all other feature extraction techniques. Whereas, IG measures how much "information" a feature provides about a particular class. There is a possibility that one feature is ranked higher by a feature selection methods and the same feature gets lower rank using another feature selection technique. Utilizing multiple feature selection methods enable to select the top tanked features based on the majority vote.

[3] https://www.nltk.org/

4.3. Classification

To identify the dominant emotion in an email, three classifiers are utilized in this work, namely Artificial Neural Network (ANN), Support Vector Machines (SVM) and Random Forest (RF). The *Enron8715*, *Emotion616*, and *Emotion135106* datasets are divided into 70:30 ratio. Seventy percent of the data is utilized for training and 30% for the testing of the classifiers. The performance of these classifiers is then evaluated using F1 score, precision, recall, and confusion matrix.

ANN: It is inspired by human brain system which consists of billions of neurons connected to each other to process and pass signals to the next connected neurons. ANNs are the networks of the artificial neurons (nodes) connected to each other. These are used in pattern recognition for supervised learning. The underlying concept of ANNs is to develop a connected network of artificial neuron so that it could learn from existing examples and predict using the previously learned data. The connected neurons can transmit signals to each other. An artificial neuron collects information from other neurons, processes it and passes it to the next neuron for further processing. ANNs are composed of the input neurons, hidden layers (optionally and problem dependent), and output neurons. The number of input neurons, hidden layers, and output neurons is problem specific. Here, the number of input neurons is equal to the number of features and the output neuron count is equal to the number of emotions, i.e., six.

SVM: The SVM is a supervised learning model used for classification. It is commonly applied to linearly separable data, however, it can also be used for non-linear classification in a high dimensional feature space. It creates a set of hyperplanes (decision planes) in a high dimensional space to classify the data. The advantage of SVM is its effectiveness in high dimensional space, its versatility (different kernel functions can be used and also the custom kernels), and also memory efficiency. However, if the number of features in larger as compared to the number of samples, overfitting may occur.

RF: Random forest is also a supervised classification method. As the name specifies, it constructs a bag of decision trees somehow random and merges them together to improve the overall result. The concept is called "bagging" method. The RF can be used for classification as well as regression problems. In classification, RF looks for the random subset of features instead of searching one best feature. It creates the diversity in RF which results in a more stable prediction. Since RF is a set of decision trees termed as ensemble method, it results in better prediction instead of a single decision tree. Ensemble learning algorithms predict on the basis of the aggregate decision of multiple predictors.

4.4. Dataset and results validation

The *Enron8715* is manually labeled and the *Emotion616* data is created using induced stimulus. To validate the existence of both positive and negative emotions, clustering is employed in this work. For this, initially, the complete data is categorized into positive or negative emotions using a benchmark tool SentiWordNet [42]. SentiWordNet is a publicly available lexicon-based resource which can be implemented using Java/Python. It measures the polarity of the words and categorizes them as *positive* or *negative*, objectivity. *Neutral*, *happy*, and *positively surprised* are termed as *positive* while *sad*, *angry* and *negatively surprised* are

included in *negative* emotion category. Afterwards, using SentiWordNet, the polarity of email is identified. Later, three clustering algorithms, namely, *k*-means, *k*-medoids, and fuzzy c-means are applied. The results of the clustering procedure are evaluated using four validity indices, namely, Davies Bouldin Index (DBI), gap clustering criterion, Silhouette Coefficient (SC), and Dunn Index (DI). Fig. 3 shows the working of the proposed work using a flowchart. The framework is mathematically shown below.

$$T = \{t : t \in \textit{the email text}\}$$
$$E = \{e : e \in \textit{emotions in the email text}\}$$
$$D = \{d : d \in \textit{Dominant in the email}\}, \quad \text{then}$$
$$T \times E \rightarrow D = \langle t, e, d \rangle \tag{1}$$

The email text dataset T consist of various instances, where each instance can be represented by the features (F),

$$T = \{t_1, t_2, t_3, \ldots, t_N\}$$
$$F = \{f_1, f_2, f_3, \ldots, f_n\}$$

The emotion set E is represented as,

$$E = \{e_1, e_2, e_3, e_4, e_5, e_6\}$$

Similarly, for the dominant emotion we have the set D,

$$D = \{d_1, d_2, d_3, d_4, d_5, d_6\}$$

This makes $N \times 6 \times 6$ possible combinations for each instance of T, E and D.

$$T \times E \times D = Z^a \cup Z^b \tag{2}$$

where Z^a *and* Z^b are two disjoint set

$$Z^a = d_i \times t_j \rightarrow v_k \tag{3}$$
$$Z^b = d_i \times t_j \neq v_k \tag{4}$$

where $i = 1, 2, 3 \ldots N, j = 1, 2, 3, \ldots, 6, M, k = 1, 2, 3, \ldots, 6.$

T represents the set of emails where each t_i has its own domain, E is a set of emotions and D is the set of dominant emotion. Later, a relation is found between T and E using a suitable mapping in set D given by Eq. (1).

5. Experimental results

This section lists the conducted experiments and their results. Emotions are intermixed feelings and to verify the presence of a specific emotion in the text, the selection of features is quite challenging. Results on the six emotions are obtained using three feature sets, i.e., all features, top features, and bottom features. To extract human emotion from email's in-text data, three classifications are evaluated, namely, Artificial Neural Network (ANN), Support Vector Machines (SVM), and Random Forest (RF). This section also contains the results of clustering for the three datasets to evaluate the distribution of the six emotions.

5.1. Parameter settings

All experiments are conducted on a computer with Intel® Core™ i5 processor having 8 GB RAM. The operating system was Microsoft Windows 10 and Python is used to code the proposed approach. For the ANN classifier, the number of input and output neurons are varied according to the specific experiment. For example, combinations of 6 input neurons/6 output neurons, 6 input neurons/2 output neurons, 14 input neurons/6 output neurons, 14 input neurons/2 output neurons, and 2 input neurons/2 output neurons are used. General parameters which remain constant are listed in Table 4.

Table 4
Parameter settings.

ANN	No of layers	3
	Activation function	Sigmoid
	Maximum Iterations	1000
	No of hidden layers	1
	No of neurons in the hidden layer	5
	Initial weights	random
	Initial learning rate	0.5
	Alpha	0.1
	Beta	0.999
SVM	Type of SVM	Polynomial
	Degree of Polynomial	5
	Stopping criteria	1.00E−03
RF	Criterion	Gini
	Number of estimators	11

Table 5
Features categorization based on PCA, IG, and MI.

Feature id	Name	Category
1	Verbs count	Top
2	Adjectives count	
3	Adverbs count	
4	Conjunction count	
5	Noun count	
6	Verb	Bottom
7	Adjective	
8	Adverb	
9	Conjunction	
10	Noun	
11	"?"	
12	"!"	
13	Positive words	
14	Negative words	

Table 6
Max. accuracy for one vs all emotions using 14 features.

Emotion label	Accuracy
Neutral	88%
Happy	84%
Sad	95%
Angry	94%
Positively Surprised	96%
Negatively Surprised	97%

5.2. Feature selection

In order to identify the key features that can assist this work in better classification, there feature selection techniques are utilized, namely, PCA, IG, and MI. Each of these techniques is executed over the *Enron8715* and *Emotion616* datasets separately. These methods are not executed over the third dataset as it is the combination of the first two, therefore the redundancy is avoided. The feature selection techniques assign a score to each of the features. Features having higher scores are considered better than those with lower scores in separating the data. Table 5 lists the results of this experiment. Where all the feature selection techniques almost have a consensus, with an exception of a few instances, over the top and bottom ranked features. According to these results, the features: verb count, adjective count, adverb count, conjunction count, and noun count are the top attributes. Whereas, the Boolean values of verbs, adjectives, adverbs, conjunction and nouns, count of the punctuation mark "?", count of punctuation mark "!", number of positive words, and number of negative words are among the bottom features.

5.3. Emotion recognition

This section contains experimental results of emotion recognition on the three datasets. The proposed framework has been

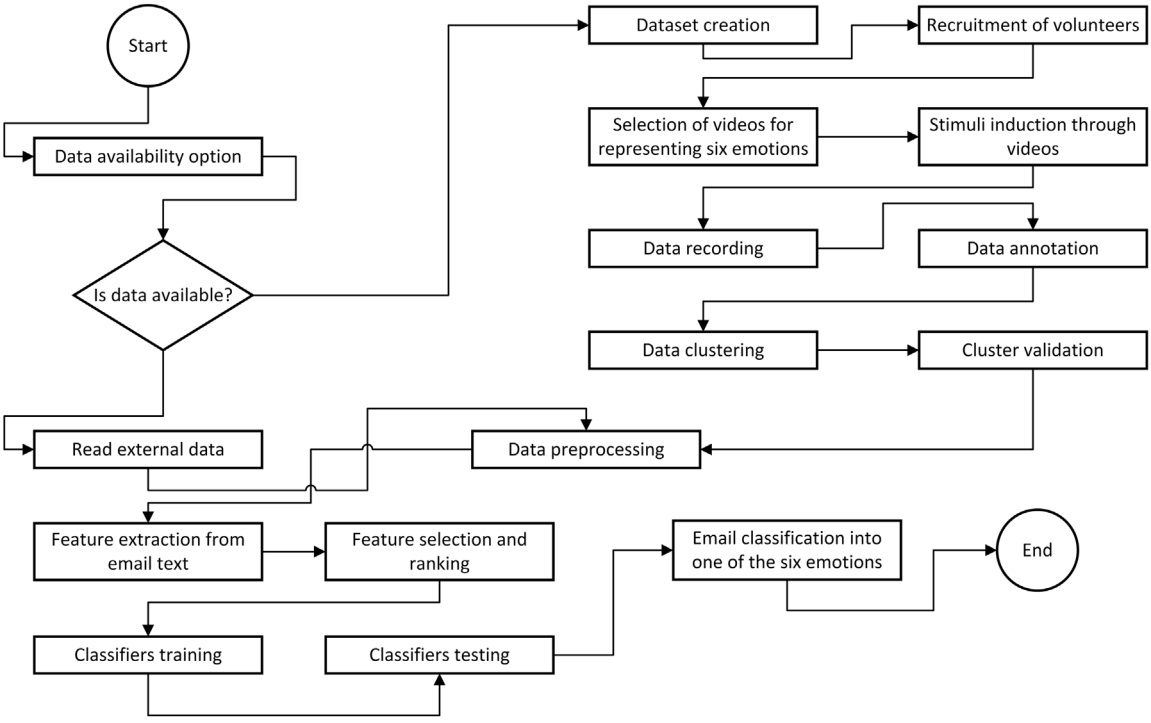

Fig. 3. Proposed work flowchart.

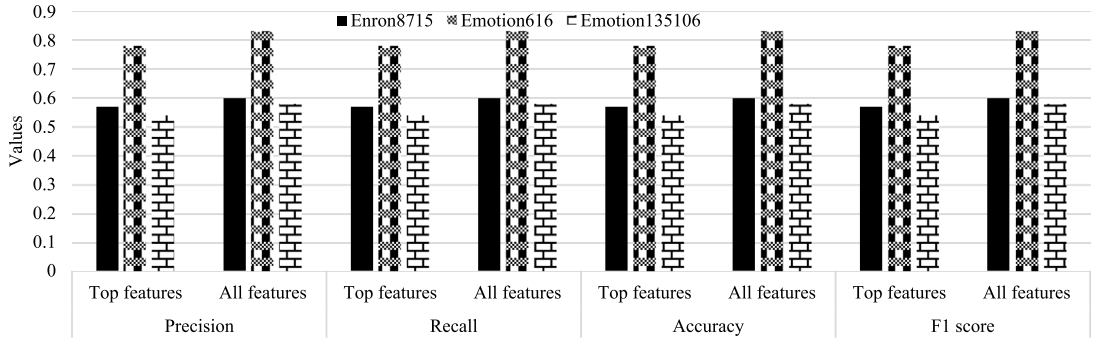

Fig. 4. Performance of the SVM classifier over the three datasets.

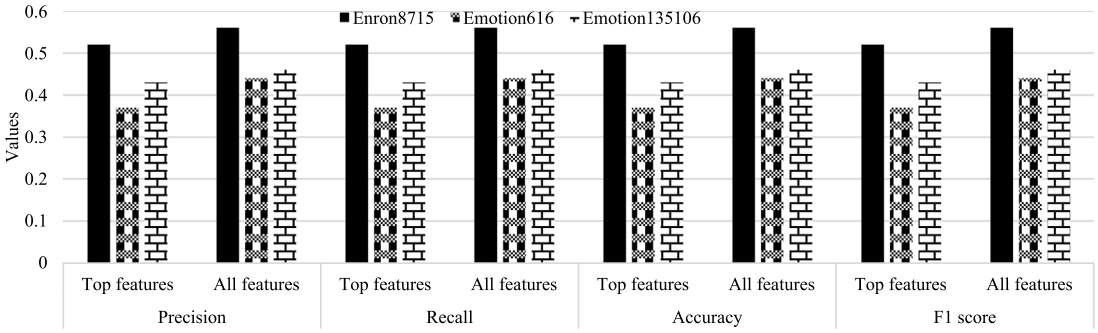

Fig. 5. Performance of the ANN classifier over the three datasets.

evaluated on three datasets, i.e., *Enron8715*, *Emotion616*, and *Emotion135106*. The *Enron8715* and *Emotion616* datasets are the unique datasets, whereas the third one is only the combination of these two. The third dataset is created to evaluate the performance of the proposed work on a relatively bigger data just like the work in [38] does it by using *StarCraft* and *World of Warcraft* datasets in addition to their own data for personality prediction. The work in [38] used additional datasets for an unsupervised learning task, therefore there was no need for combining the datasets. Whereas, the present work is addressing a supervised learning problem due to which the first two datasets are merged to create a third and larger one. This has the benefit of obtaining evaluations on a bigger dataset. Fig. 4 shows the classification results of SVM using top-ranked features and all features. The

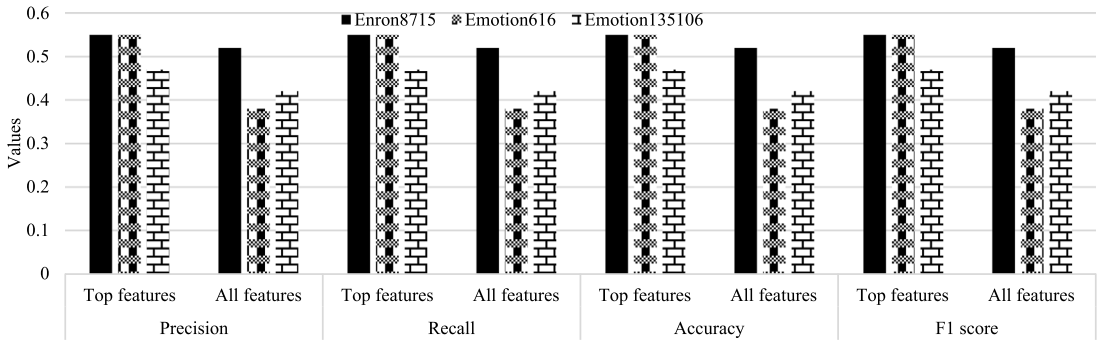

Fig. 6. Performance of the RF classifier over the three datasets.

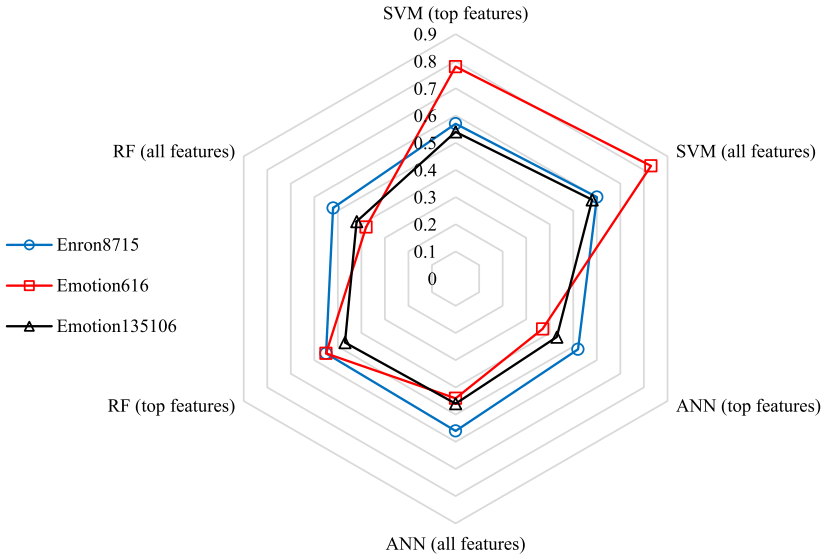

Fig. 7. Comparative performance of the three classifiers.

SVM has 83% accuracy on the *Emotion616* dataset using all features and 78% using only the top features. As the distribution of emotion label is quite even in *Emotion616*, therefore, SVM has a better classification accuracy in this case. Fig. 4 also shows other performance evaluation metrics of classification, i.e., precision, recall, and F1 score. The ANN is implemented using two combinations of neurons. The first combination has six input neurons, 5 hidden layer neurons, and 6 output neurons. Whereas, the second one has 14 input neurons, 5 hidden layer neurons, and 6 output neurons. Fig. 5 shows accuracy results of ANN using top features and all features. The ANN gives a maximum of 56% accuracy using all features on the *Enron8715* dataset and the average accuracy of the ANN is 55%. The results of the random forest classifier are shown in Fig. 6. The random forest attains only 55% accuracy using top features. Fig. 7 shows a comparative performance of the three classifiers. It can be seen from the figure that when the proposed framework incorporates SVM in its prediction phase, a maximum accuracy of 83% in achieved for the *Emotion616* dataset. The average accuracy for all datasets using SVM is 67%. Whereas, the ANN performs the second best with an accuracy of 56% on the *Enron8715* dataset. It has an average accuracy of 48% for all datasets, which is fairly low. Although the RF perform towards a lower side as compared to the other two classifiers, however, it achieves better performance with the top-ranked features, i.e., 55% on the *Emotion616* data as compared to its performance on all features. Overall, majority of the classifiers perform better on the *Emotion616* data in comparison to the other two datasets. A reason to this can be it being balanced. Additionally, the *Emotion616* dataset is specifically creating for identifying human emotion hidden in the email text. This is done by inducing six basic emotions in participants while recording the data. This enables the data to have maximum possible keywords against each of the emotion under consideration. Whereas, majority of the emails in the *Enron8715* data are annotated as *neutral* or *happy* resulting in relatively low performance of the classifiers on this dataset.

5.4. Emotion recognition for specific emotion vs all

An experiment has been performed to evaluate the classification accuracy by predicting one class versus all. For example, the grouping of emotion labels for *neutral* emotion is, *neutral* emotion versus other five emotions. All the other emotions are assigned the same label in this case. This is repeated for all six emotions. Later, supervised machine learning techniques are implemented using ANN, SVM, and RF algorithms. Here, the maximum accuracy of 97% is observed for *negatively surprised* emotion. Table 6 shows the accuracy of classification results for all emotions. Fig. 8 show the emotion classification accuracy of the six emotions on *Enron8715*, *Emotion616*, and *Emotion135106* datasets.

5.5. Sentiment classification using SentiWordNet

In addition to the identification of dominant emotion from the email text, experiments are performed using the proposed framework to identify prominent sentiment. The previous experiments demonstrated the utility of the proposed work for

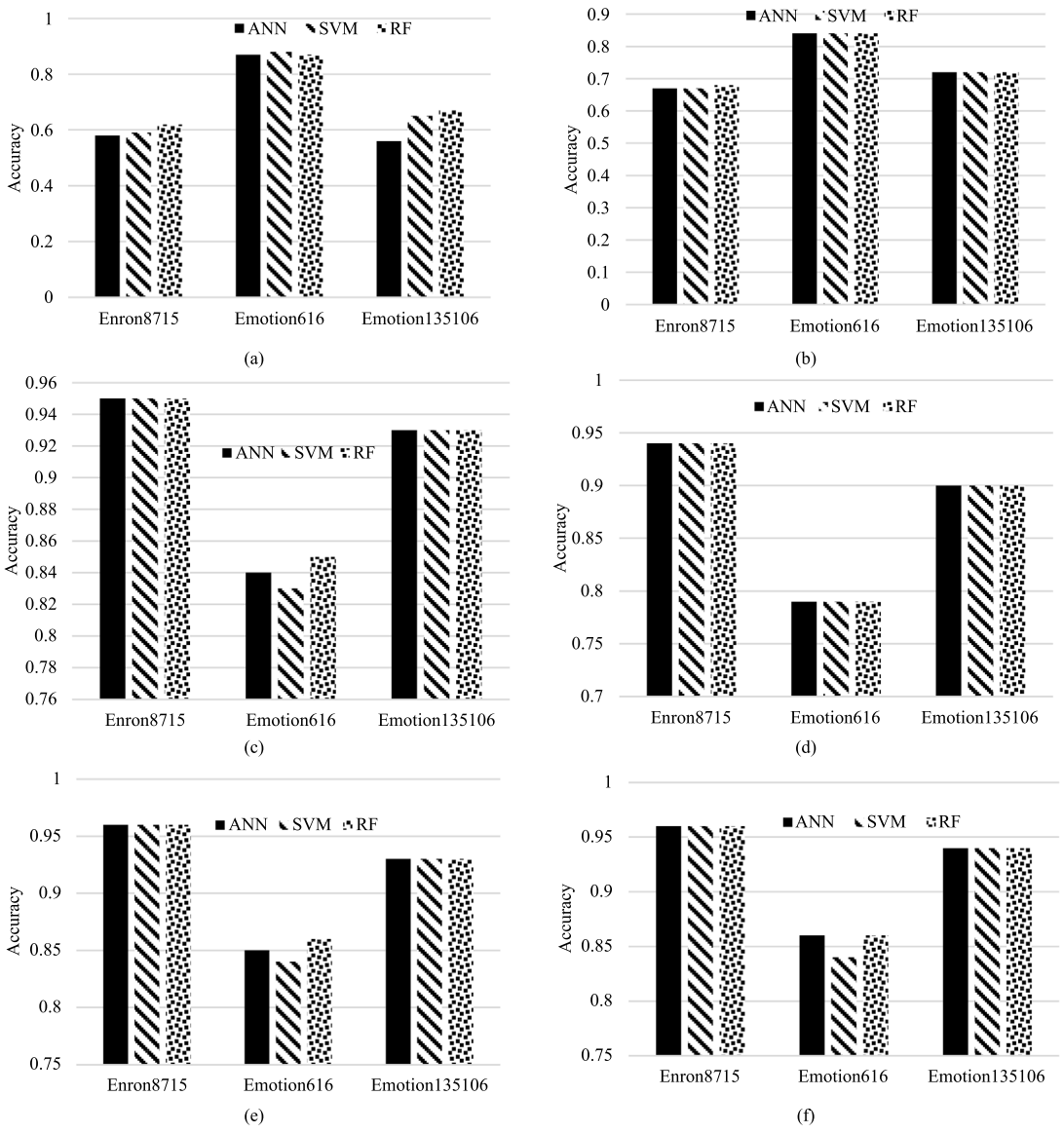

Fig. 8. Performance of the three classifiers using one vs. others strategy. (a) Neutral vs. others (b) Happy vs. others (c) Sad vs. others (d) Angry vs. others (e) Positively surprised vs. others (f) Negatively surprised vs. others.

emotion recognition, this experiment shows the generality of the present work by predicting sentiments as well. Emotion identifying is a high dimensional problem having six labels, whereas, the sentiment analysis is a bipolar task (excluding the neutral class). For this experiment, the first undertaking is to convert the six classed data into a binary class representation. SentiWordNet is a resource used to extract opinion and sentiments in the text analysis domain. It is available for Python to use with the Natural Language ToolKit (NLTK). For the transformation of six classes into two, features 13 and 14 are used in this work, i.e., counts of positive and negative words are extracted from the two original email datasets through the SentiWordNet. Later, the class labels are divided into two groups; *positive* and *negative*. The *neutral*, *happy*, and *positively surprised* emotions are labeled as *positive* group. Whereas, the *negative* group is formed using *sad*, *angry*, and *negatively surprised* emotions. Positive word count contributes to the prediction of the *positive* class and *negative* word count contributes to the *negative* class. Once the binary class representation of the datasets is obtained, it is verified through clustering before conducting any further experiments. The clustering results are evaluated using the Davies Bouldin Index (DBI), Gap Evaluation (GE), and Silhouette Coefficient (SC). The smallest value of DBI, the highest value of GE and the highest value of SC indicates the optimum clustering formations. To avoid redundancy, the third dataset is not used for validation through clustering because it is a merger of the first two datasets. Fig. 9(a) and (b) shows the clustering results using binary labeled *Enron8715* and *Emotion616* datasets. Where, it can be observed that the cluster validity indices suggest better cluster formation for two groups formed using the abovementioned procedure. After confirming the proper transformation of the datasets, the proposed framework is used for sentiment prediction.

The obtained results are evaluated using precision, recall, accuracy, and F1 score. Classification results on *Enron8715* show 86% accuracy for ANN, SVM, and RF as indicted in Table 7. It shows the results of ANN, SVM, and RF on *Emotion616* dataset having 55%, 59%, and 63% accuracy, respectively. Table 7 shows 75%, 76%, and 76% accuracy using ANN, SVM, and RF on *Emotion135106* dataset. The key performance indicators of these classifiers are listed in Table 8. It can be seen from the results that the RF when incorporated in the proposed framework for classification achieves the highest average accuracy of 75%. Whereas, based

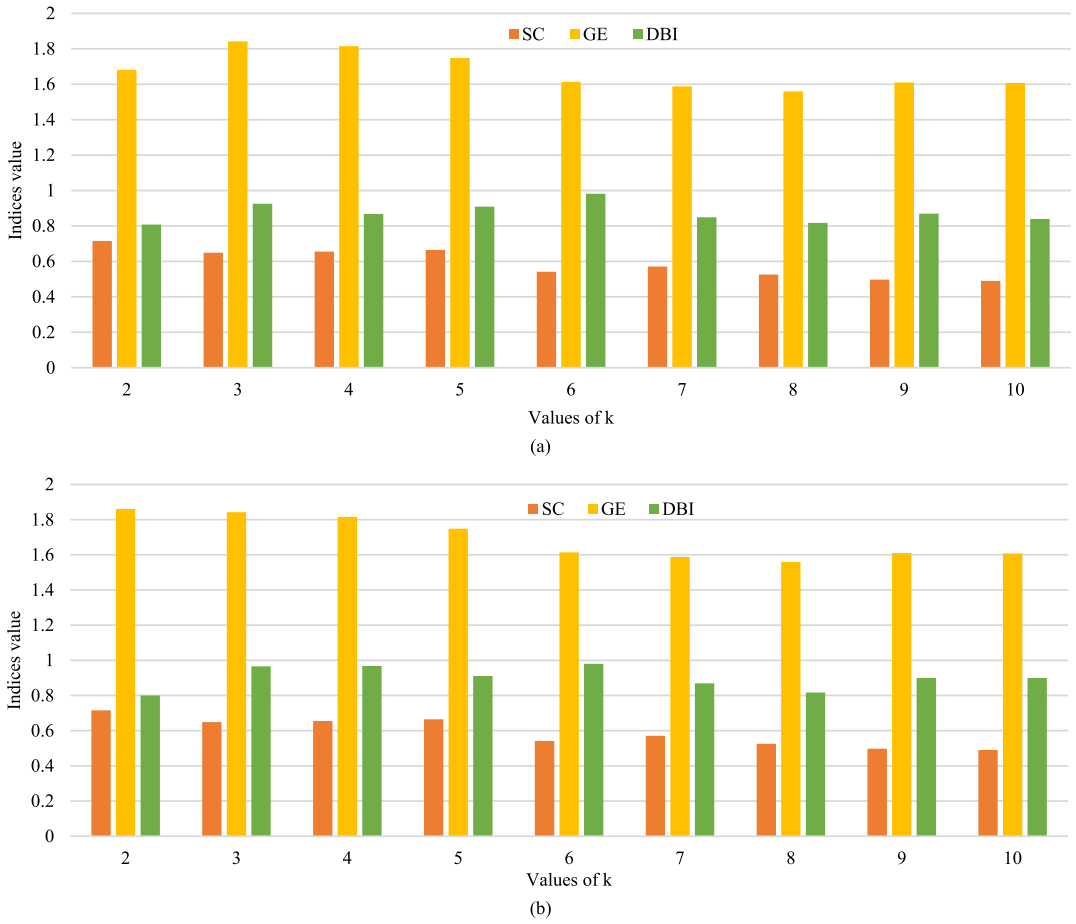

Fig. 9. Cluster validity indices for *k*-means (a) *Enron8715* dataset (b) *Emotion616* dataset.

Table 7
Sentiment classification results using SentiwordNet on the three datasets.

Dataset	Classifier	Precision	Recall	Accuracy	F1 score
Enron8715	ANN	0.86	0.86	0.86	0.86
	SVM	0.86	0.86	0.86	0.86
	RF	0.88	0.86	0.86	0.86
Emotion616	ANN	0.55	0.55	0.55	0.55
	SVM	0.59	0.59	0.59	0.59
	RF	0.63	0.63	0.63	0.63
Emotion135106	ANN	0.75	0.75	0.75	0.75
	SVM	0.76	0.76	0.76	0.76
	RF	0.76	0.76	0.76	0.76

Table 8
Sentiment classification results using SentiwordNet.

Dataset	Classifier	Precision	Recall	Accuracy	F1 score
Enron8715	ANN	0.86	0.86	0.86	0.86
	SVM	0.86	0.86	0.86	0.86
	RF	0.88	0.86	0.86	0.86
Emotion616	ANN	0.55	0.55	0.55	0.55
	SVM	0.59	0.59	0.59	0.59
	RF	0.63	0.63	0.63	0.63
Emotion135106	ANN	0.75	0.75	0.75	0.75
	SVM	0.76	0.76	0.76	0.76
	RF	0.76	0.76	0.76	0.76

on the average accuracy one all datasets, ANN performs the least with an average accuracy of 72%. Observing the results of this experiment separately on each dataset, all classifiers achieve maximum accuracy on the *Enron8715* dataset. Whereas, the RF classifier performs better than others by achieving maximum accuracy on all three datasets. Additionally, similar trend can be observed for RF based on precision, recall, and F1 score metrics. ANN in this case has performed the least on all three datasets.

5.6. Sentiment classification using all features and top features

As mentioned earlier, sentiment classification can involve two types of grouping. First one is the grouping of emotions as *positive* or *negative*. Whereas, the second is the cataloging of emotions into *neutral*, *positive*, or *negative* categories. Here, the first group of categorization is used, i.e., *positive* and *negative*. To classify an email text, both groups (top features and all features) are applied. The classification results of top features are shown in Fig. 10 where up to 86% accuracy is achieved through this classification approach. The ANN gives the best accuracy in comparison to SVM and RF. Fig. 10 shows the results of ANN, SVM, and RF using all features. The highest accuracy is still 86% by ANN, however, the performance of RF is decreased to 68% on the *Emotion135106* dataset.

5.7. Clustering

The distribution of feature values for the six emotions is evaluated by means of clustering, an unsupervised learning method. The *k*-means, *k*-medoids, and fuzzy c-means clustering algorithms are applied and their results are evaluated using DI. All three groups of features, i.e., top features, bottom features, and all features are utilized to perform the clustering procedure separately. Fig. 11 shows the clustering results obtained using

fuzzy c-means algorithm implemented on the *Emotion616* dataset which gives better cluster validity index value at $k = 6$. Therefore, supporting the existence of six different categories (emotions in this case) in the underlying dataset. A similar pattern is observed for all set of features. The *k*-means clustering results are shown in Fig. 11. DI value on *k*-means clustering results gives highest cluster validity indices values for six clusters. Bottom features set do not perform well in the case of *k*-means clustering. The third clustering algorithm, i.e., *k*-medoids also has optimum results for six number of clusters. Overall, the feature set consisting of top features obtains best clustering formation for the cluster validity measure of DI at $k = 6$.

6. Discussion

This work presented an approach to identify the dominant emotion hidden in an email using its in-text data. For this purpose, the Enron email dataset containing 0.5 million emails of 150 organizational officers was utilized. However, the dataset was not labeled therefore the labeling was done as part of the current work. In addition to this, a second dataset was created utilizing 60 participants. For this, stimuli were induced to create an environment specific to each of the six emotions. Later, the participants were instructed to write emails on a specific topic while they were under the influence of the induced emotion. This dataset consisted of around 339 emails. The third dataset was the combined form of the first two datasets. Next, these datasets were tokenized into separate words to find frequencies of each of the words and discard any redundant portions. The stop words were removed from the list of tokens and lemmatization of the remaining tokens was done. This process utilized PoS and context for conversion of a word into the root word. Later, 14 features, namely number of verbs, number of adjectives, number of adverbs, number of conjunction, number of nouns, five features containing Boolean values of verbs, adjectives, adverbs, conjunction and nouns, count of punctuation mark "?", count of punctuation mark "!", number of positive words, and number of negative words were extracted from the three datasets to train the classifiers. Three feature selection techniques were utilized to identify the most suitable features in extracting the dominant emotion from the email text. This resulted in top features (i.e., verb count, adjective count, adverb count, conjunction count, and noun count), bottom features (i.e., the Boolean values of verbs, adjectives, adverbs, conjunction and nouns, count of the punctuation mark "?", count of punctuation mark "!", number of positive words, and number of negative words), and all features. Three classifiers, i.e., ANN, SVM, and RF were utilized to learn the patterns of an email against each of the six emotion. Classification results of *Enron8715* gave 86% accuracy for ANN, SVM and RF. The ANN, SVM, and RF on *Emotion616* dataset had 55%, 59%, and 63% accuracy, respectively. Whereas, on *Emotion135106* dataset ANN, SVM, and RF had 75%, 76%, and 76% accuracy, respectively.

6.1. Results analysis

Emotion is a psychological state that involves individual experience, physiological response, and behavioral reaction. Emotions do not exist in isolation, but are a mixture of different overlapping feelings. Therefore, to recognize a specific emotion contained in text as compared to just evaluating the polarity (*positive* or *negative*) is quite challenging. This research work presented a number of experiments for emotion recognition using a combination of emotion labels. A number of different classifiers and clustering techniques were applied. The classification results of six emotion classes showed up to 83% accuracy through SVM while the ANN and RF gave approximately 56% and 55% accuracy, respectively. However, when the emotion class labels were combined into two groups, i.e., *positive* and *negative* the accuracy increased up to 86%. Similarly, when emotion class labels were grouped as one emotion label against all others, the accuracy ranged between 84% to 97%. Utilization of the SentiWordNet tool showed a maximum of 86% accuracy to predict *positive* and *negative* sentiments in an email. Different combinations of features were also used for both classification and clustering. The last two features, i.e., the count of positive words and count of negative words showed 86% classification accuracy for *positive* and *negative* sentiments. The group of 14 features performed well as compared to the group of top features alone and the two features (count of positive words and count of negative words). Clustering techniques were also implemented using the set of all features and set of two features. To check the distribution of six different types of emotions contained in the datasets, three clustering techniques were applied, i.e., *k*-means, *k*-medoids, and fuzzy c-means. On one hand, the set of all features gave better clustering validation value for 6 number of clusters. On the other hand, the set of two features gave optimum value for two clusters using multiple cluster validation metrics.

6.2. Implications of the proposed work

Email is a professional way of communication which is used worldwide by organizations and individuals. Intentionally or unintentionally, the sender of an email always sends so with some type of emotion. Email contents are at times meager in nature. Business use of email has its advantages as well as a very few limitations. Advantages include flexibility, efficiency, and asynchronous facility. Whereas, the limitations include misinterpretation of an email data due to short text and other relational and informational effects. The receiver of an email mostly interprets the email based on her point of view which may lead to miscommunication on the personal or organizational level. This miscommunication can influence the business intelligence and decision making of the organization or relationship issues at a personal level. The present proposal enables the email service providers to extract the dominant emotion hidden in an email to interpret it in a proper way. This will enable the users to avoid any misunderstanding during personal communication via emails and for organizations to assist in appropriate decision making. Additionally, various vendors can utilize this feature to interpret the true emotion of their clients' queries while responding via emails.

6.3. Comparison

This section presents a comparison of the proposed approach with the closely related past contributions on emotion recognition. Both, quantitative and qualitative comparisons have been made here. For the quantitative comparison, two newer methods are opted for, namely, Bostan et al. [43] and Tripathi et al. [5]. Whereas, the rest of the methods (including those for sentiment analysis) are compared utilizing qualitative approach.

Quantitative comparison: The proposed approach is compared with Bostan et al. [43] and Tripathi et al. [5]. Bostan et al. [43] aggregates multiple emotion classification datasets in a common file format using same annotation schema. Additionally, they use maximum entropy classifiers for the identification of emotion in a dataset having six class labels. All comparisons are made here on the same datasets for the sake of fairness. Fig. 12 shows the best case of the three competing methods. It can be seen that the proposed framework performs better than others in the majority of the cases. The average accuracy of the proposed approach is 67% on the three datasets and it obtains an accuracy of 83% on

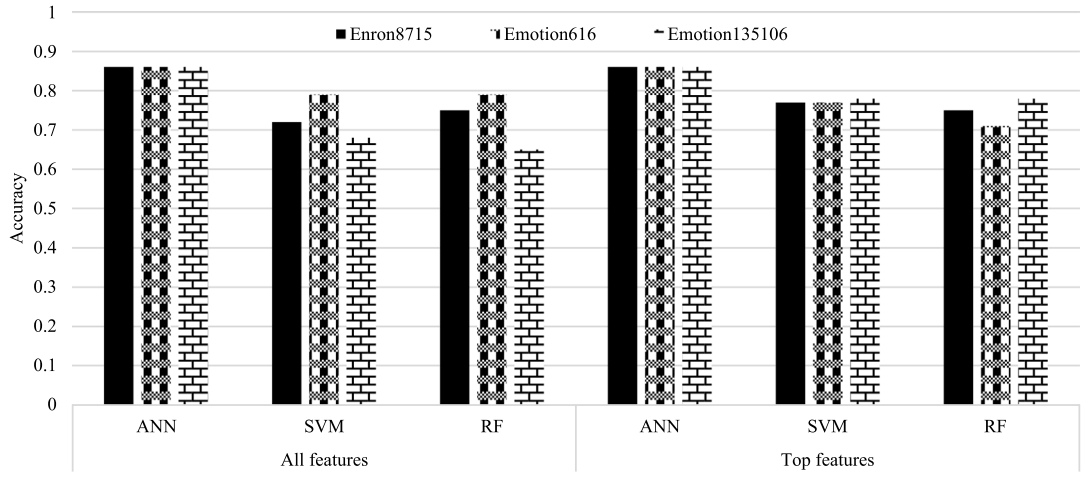

Fig. 10. Sentiment classification performance of the three classifiers into two categories using top and all features.

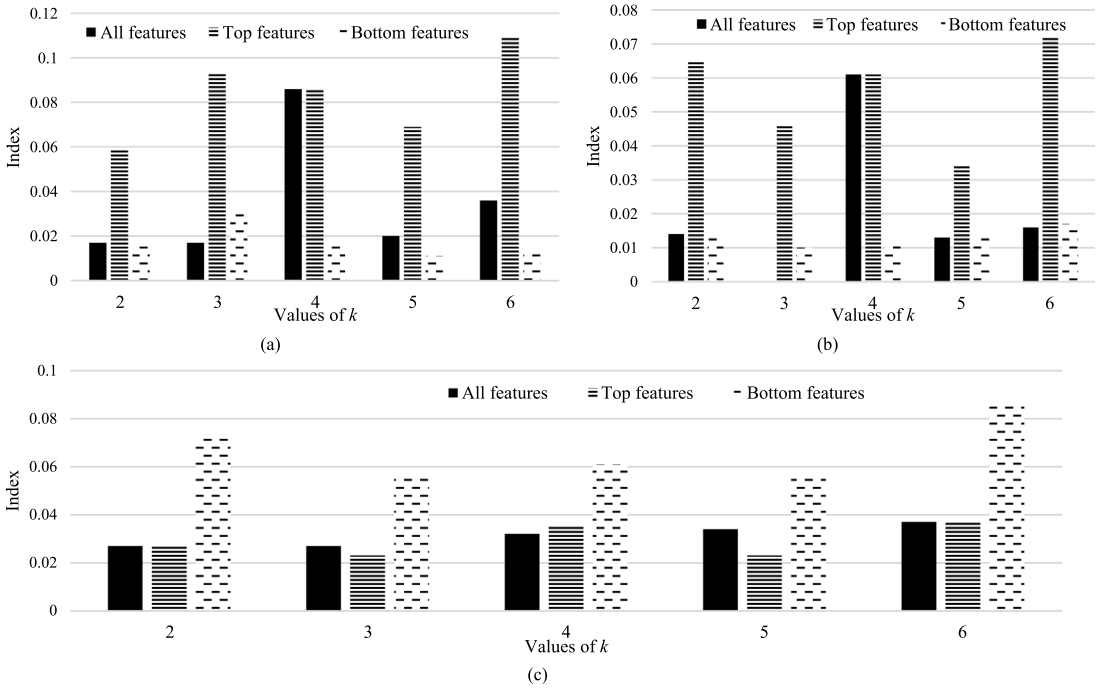

Fig. 11. Cluster validity index DI values (a) *k*-means clustering (b) *k*-medoids clustering (c) Fuzzy C-means Clustering.

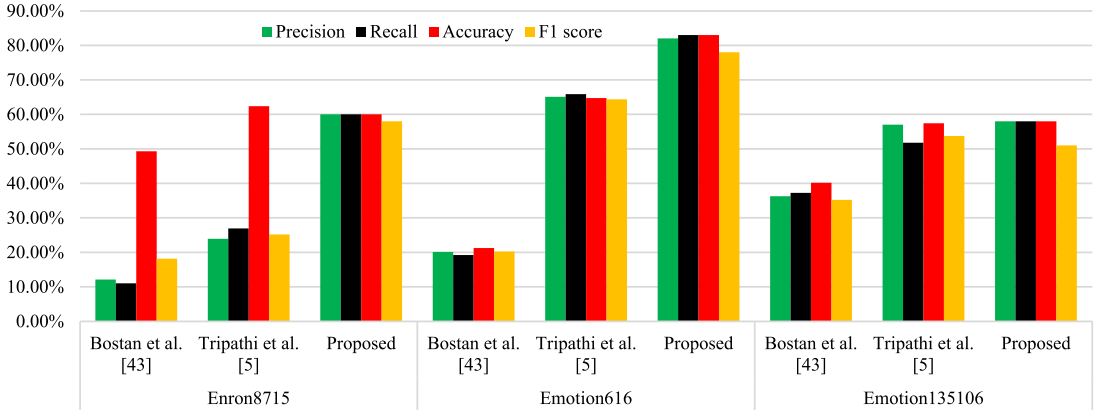

Fig. 12. Quantitative comparison of the proposed approach and two state-of-the-art methods.

Table 9
Qualitative comparison between the proposed framework and past works.

Methods	Reported accuracy (best case)	Emotions count	Technique	Flexibility
Proposed	83.00%	6	ML-based framework	✓
Kratzwald et al. [17]	64.70%	4	RANN	○
Tripathi et al. [5]	71.04%	4	ML+DL	○
Matsumoto et al. [18]	36.00%	8	Semantic similarity	○

Table 10
Key strengths and limitations of the proposed approach.

Strengths	Limitations
Rich dataset with ground truth available	Role of gender overlooked
Better accuracy	Gender imbalance in the dataset
Rigorous analysis	

the primary dataset, i.e., *Emotion616*. Tripathi et al. [5] has an accuracy average of 61% on the three datasets and it obtains an accuracy of 64.71% on the primary dataset. Whereas, the Klinger et al. method has an average accuracy of 37% and 21.23% accuracy on the primary dataset. The approach proposed in Tripathi et al. performs slightly better than the current work on *Enron8715* dataset only with a difference of 2%.

Qualitative comparison: According to the closely related past works on identification of emotion from text, the classification results of [17] gave 64.7% accuracy using the SVM as a classifier. Similarly, Tripathi et al.'s model [5] had a maximum of 64.78% accuracy for emotion detection from text-based data. Other previous emotion recognition work include [18] having 36% accuracy for eight emotion labels. However, while considering the classification results of sentiment analysis only, an accuracy of almost 99% is achieved in the past works, for example, Ren et al. [24] reported 99% accuracy and Xia et al. [25] mentioned 90% accuracy for textual data. Sentiment analysis of email textual data by categorizing into one of the three classes, i.e., *positive*, *negative*, and *neutral* showed 98% accurate results [10]. All this suggests comparable performance of the current proposal with the related past works. Table 9 lists the qualitative comparison of the proposed work and past contributions for identifying the emotion.

This proposal performed a detailed set of experiments by utilizing three feature selection techniques, 14 features, three classifiers, three clustering techniques, and four cluster validity indices. The artificial neural network and random forest did not perform well as a classifier for six emotions. Whereas, the support vector machines gave appropriate results with 83% accuracy for six emotions. The Enron dataset contains approximately 0.5M emails while here 1000 emails were selected for experiments. These emails were first annotated by three annotators followed by its utilization in the experiments. This process required careful annotation and this is the reason for the limited number of emails used here. The other two datasets had 339 and 1349 samples of labeled emails. Experiments can be done using a larger number of labeled samples in the future. According to the literature review, there are a number of annotation methods, for example, single class labeling (used in this work), multiclass labeling in which annotators can assign more than one output class labels to the data, and scoring annotation methods. Experiments can be performed to evaluate various annotation methods and to see their effectiveness in identifying the emotions. Moreover, here SentiWordNet was implemented for sentiment analysis to predict only *positive* and *negative* emotion and it showed better results. In the future, SentiWordNet can be explored further to detect specific emotion instead of only examining the polarity of email data. As mentioned above, the artificial neural network and random forest did not perform well for emotion recognition in this work in comparison to the SVM. To address this, different hybrid approaches can be implemented for classification may combine existing supervised machine learning algorithms [44,45]. Other than the novelty and strengths of this proposal, there are a few limitations. This work has overlooked the role of gender in identification of a particular emotion in the email. Gender plays an important role in studies related to the identification of emotion since a few emotions are very strong for a specific gender. This study does not address the influence of various cultures on emotion identification as well. The second dataset i.e., *Emotion616* involved 61 volunteers, most of these were teenagers. Table 10 lists the key strengths and limitations of this proposal.

7. Conclusion

Emotions are a basic part of human nature. This work presented a supervised machine learning-based framework for emotion recognition from email text. The proposed approach was also provided with a labeled and balanced dataset for evaluation. For this, 60 volunteers were engaged and a stimulus was induced for each of the six emotions, namely, *neutral*, *happy*, *sad*, *angry*, *positively surprised*, and *negatively surprised*. Later, the participants were asked to write an email on a specific subject for each of the emotion types. This provided 339 emails after discarding low quality samples. In addition to this, a publicly available dataset was also used in this work after proper annotation. The proposed framework was incorporated with three feature selection methods, three classification methods, two clustering techniques, and multiple cluster validity indices. Experiments were performed by extracting 14 features from the datasets. Three feature selection techniques, i.e., Principal Component Analysis (PCA), Mutual Information (MI), and Information Gain (IG) were used to identify the optimum features for classification. Afterwards, three classifiers, namely, Artificial Neural Network (ANN), Random Forest (RF), and Support Vector Machines (SVM), were employed to predict the emotions hidden in the emails. Classifiers' performance was evaluated using F1 score, precision, accuracy, and recall. Based on the feature selection results, experiments were conducted on all three datasets by vertically partitioning them into all features, top features, and bottom features. During experiments, 83% accuracy was achieved for the proposed work to predict an emotion. Qualitative and quantitative comparisons of the proposed work were also made with two state-of-the-art methods. The obtained results suggested better performance of the current work. This proposal enables the email service providers to have an add-on for identifying the dominant emotion hidden in an email to interpret it properly. This will enable the users to avoid any misunderstanding during personal communication via emails and for organizations to assist in appropriate decision making. Additionally, various vendors can utilize this feature to interpret the true emotion of their clients' queries while responding through emails.

CRediT authorship contribution statement

Zahid Halim: Conceptualization, Methodology, Software, Writing - review & editing. **Mehwish Waqar:** Data curation, Writing - original draft, Software. **Madiha Tahir:** Visualization, Investigation, Software.

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.

Acknowledgment

The authors are indebted to the editor and anonymous reviewers for their helpful comments and suggestions. The authors wish to thank GIK Institute for providing research facilities. This work was sponsored by the GIK Institute graduate research fund under GA1 scheme.

Funding

This work was sponsored by the GIK Institute graduate research fund under GA1 scheme. Grant number GCS1635.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent

Informed consent was obtained from all individual participants included in the study.

References

[1] M. Lenartowicz, Creatures of the semiosphere: A problematic third party in the 'humans plus technology'cognitive architecture of the future global superintelligence, Technol. Forecast. Soc. Change 114 (2017) 35–42.
[2] Z. Halim, M. Rehan, On identification of driving-induced stress using electroencephalogram signals: A framework based on wearable safety-critical scheme and machine learning, Inf. Fusion 53 (2020) 66–79.
[3] A.K. Przybylski, N. Weinstein, A large-scale test of the goldilocks hypothesis: quantifying the relations between digital-screen use and the mental well-being of adolescents, Psychol. Sci. 28 (2) (2017) 204–215.
[4] K.P. Seng, L.M. Ang, C.S. Ooi, A combined rule-based & machine learning audio-visual emotion recognition approach, IEEE Trans. Affect. Comput. 9 (1) (2018) 3–13.
[5] S. Tripathi, H. Beigi, Multi-modal emotion recognition on IEMOCAP dataset using deep learning, 2018, arXiv preprint arXiv:1804.05788.
[6] R. Hogenraad, Fear in the West: a sentiment analysis using a computer-readable Fear Index, in: Quality & Quantity, 2018, pp. 1–23.
[7] M. Etter, E. Colleoni, L. Illia, K. Meggiorin, A. D'Eugenio, Measuring organizational legitimacy in social media: Assessing citizens' judgments with sentiment analysis, Bus. Soc. 57 (1) (2018) 60–97.
[8] W. Zhang, W. Meng, Z. Yan-chun, Does government information release really matter in regulating contagion-evolution of negative emotion during public emergencies? From the perspective of cognitive big data analytics, Int. J. Inf. Manage. (50) (2020) 498–514.
[9] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, S. Poria, Multimodal sentiment analysis using hierarchical fusion with context modeling, Knowl.-Based Syst. 161 (1) (2018) 124–133.
[10] S. Liu, I. Lee, A hybrid sentiment analysis framework for large email data, in: International Conference on Intelligent Systems and Knowledge Engineering, ISKE, 2015, pp. 324–330.
[11] P. Ekman, Facial expressions of emotion: an old controversy and new findings, Phil. Trans. R. Soc. B 335 (1273) (1992) 63–69.
[12] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, in: Advances in Neural Information Processing Systems, 2015, pp. 649–657.
[13] R. Litman, A. Bronstein, M. Bronstein, U. Castellani, Supervised learning of bag-of-features shape descriptors using sparse coding, Comput. Graph. Forum 33 (5) (2014) 127–136.
[14] C. Dos Santos, M. Gatti, Deep convolutional neural networks for sentiment analysis of short texts, in: Proceedings of 25th International Conference on Computational Linguistics, 2014, pp. 69–78.
[15] Y. Rao, H. Xie, J. Li, F. Jin, F.L. Wang, Q. Li, Social emotion classification of short text via topic-level maximum entropy model, Inf. Manage. 53 (8) (2016) 978–986.
[16] R. Jongeling, S. Datta, A. Serebrenik, Choosing your weapons: On sentiment analysis tools for software engineering research, in: 2015 IEEE International Conference on Software Maintenance and Evolution, ICSME, 2015, pp. 531–535.
[17] B. Kratzwald, S. Ilic, M. Kraus, S. Feuerriegel, H. Prendinger, Decision support with text-based emotion recognition: Deep learning for affective computing, 2018, arXiv preprint arXiv:1803.06397.
[18] K. Matsumoto, M. Yoshida, Q. Xiao, X. Luo, K. Kita, Emotion recognition for sentences with unknown expressions based on semantic similarity by using Bag of Concepts, in: IEEE 12th International Conference on Fuzzy Systems and Knowledge Discovery, 2015, pp. 1394–1399.
[19] W. Amelia, N.U. Maulidevi, Dominant emotion recognition in short story using keyword spotting technique and learning-based method, in: International Conference On Advanced Informatics: Concepts, Theory And Application, 2016, pp. 1–6.
[20] J.M. Talarico, D. Berntsen, D.C. Rubin, Positive emotions enhance recall of peripheral details, Cogn. Emotion 23 (2) (2009) 380–398.
[21] O. Dare, C.C. Douglas, An Analysis Tool to Methodically Scan Enterprise Scale Email Systems: What's in Your Email System? in: International Symposium on Distributed Computing and Applications to Business, Engineering and Science, 2017, pp. 67–70.
[22] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Eng. J. 5 (4) (2014) 1093–1113.
[23] A. Islam, D. Inkpen, Semantic text similarity using corpus-based word similarity and string similarity, ACM Trans. Knowl. Discov. Data 2 (2) (2008) 10.
[24] F. Ren, Y. Wu, Predicting user-topic opinions in Twitter with social and topical context, IEEE Trans. Affect. Comput. 4 (4) (2013) 412–424.
[25] R. Xia, F. Xu, C. Zong, Q. Li, Y. Qi, T. Li, Dual sentiment analysis: Considering two sides of one review, IEEE Trans. Knowl. Data Eng. 27 (8) (2015) 2120–2133.
[26] X. Chen, M. Vorvoreanu, K. Madhavan, Mining social media data for understanding students' learning experiences, IEEE Trans. Learn. Technol. 7 (3) (2014) 246–259.
[27] D. Bollegala, D. Weir, J. Carroll, Cross-domain sentiment classification using a sentiment sensitive thesaurus, IEEE Trans. Knowl. Data Eng. 25 (8) (2013) 1719–1731.
[28] L. Zafar, I. Ahmed, M. Aleem, M.A. Islam, M.A. Iqbal, Analyzing adverbs impact for sentiment analysis using hadoop, in: IEEE International Conference on Emerging Technologies, 2017, pp. 1–6.
[29] L.A. Cabrera-Diego, N. Bessis, I. Korkontzelos, Classifying emotions in Stack Overflow and JIRA using a multi-label approach, Knowl.-Based Syst. 195 (2020) 05633.
[30] F. Greco, A. Polli, Emotional text mining: Customer profiling in brand management, Int. J. Inf. Manage. 51 (2020) 101934.
[31] S. Bonilla, E. Moguel, J. Garcia-Alonso, J. Berrocal, J.M. Murillo, Emotion identification with smartphones to improve the elder quality of life using facial recognition techniques, in: Exploring the Role of ICTs in Healthy Aging, 2020, pp. 178–193.
[32] X. Gao, Q. Sun, H. Xu, J. Gao, Sparse and collaborative representation based kernel pairwise linear regression for image set classification, Expert Syst. Appl. 140 (2020) 112886.
[33] J. Gao, L. Li, B. Guo, A new extendface representation method for face recognition, Neural Process. Lett. 51 (1) (2020) 473–486.
[34] T. Zhang, X. Wang, X. Xu, C.P. Chen, GCB-Net: Graph convolutional broad network and its application in emotion recognition, IEEE Trans. Affect. Comput. (2019) http://dx.doi.org/10.1109/TAFFC.2019.2937768.
[35] Z. Halim, O. Ali, G. Khan, On the efficient representation of datasets as graphs to mine maximal frequent itemsets, IEEE Trans. Knowl. Data Eng. (2019) http://dx.doi.org/10.1109/TKDE.2019.2945573.
[36] J. Gao, L. Li, A robust geometric mean-based subspace discriminant analysis feature extraction approach for image set classification, Optik 199 (2019) 163368.
[37] L. Erkoreka, I. Zamalloa, S. Rodriguez, P. Muñoz, A. Catalan, A. Arrue, M.I. Zamalloa, M.A. Gonzalez-Torres, M. Zumarraga, Genetic modulation of facial emotion recognition in borderline personality disorder, Prog. Neuro-Psychopharmacol. Biol. Psychiatry 99 (2020) 109816.
[38] Z. Halim, M. Atif, A. Rashid, C.A. Edwin, Profiling players using real-world datasets: Clustering the data and correlating the results with the big-five personality traits, IEEE Trans. Affect. Comput. 10 (4) (2019) 568–584.
[39] S. Liu, I. Lee, G. Cai, Sentiment Clustering with Topic and Temporal Information from Large Email Dataset, in: Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Posters, 2016, pp. 363–371.
[40] J. Shetty, J. Adibi, The Enron Email Dataset Database Schema and Brief Statistical Report, Vol. 4, Information sciences institute technical report, (1) University of Southern California, 2004, pp. 120–128.

[41] C.O. Alm, D. Roth, R. Sproat, Emotions from text: machine learning for text-based emotion prediction, in: Proceedings of the conference on human language technology and empirical methods in natural language processing, 2005, pp. 579–586.
[42] S. Baccianella, A. Esuli, F. Sebastiani, Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining, In Lrec 10 (2010) 2200–2204.
[43] L. Bostan, R. Klinger, An analysis of annotated corpora for emotion classification in text, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 2104-2119.
[44] T. Muhammad, Z. Halim, Employing artificial neural networks for constructing metadata-based model to automatically select an appropriate data visualization technique, Appl. Soft Comput. 49 (2016) 365–384.
[45] Uzma, Z. Halim, Optimizing the DNA fragment assembly using metaheuristic-based overlap layout consensus approach, Appl. Soft Comput. 92 (2020) 106256.