Journal of Computational Science 21 (2017) 316–326

Contents lists available at ScienceDirect

Journal of Computational Science

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

Extraction of emotions from multilingual text using intelligent text processing and computational linguistics

Vinay Kumar Jain[a,*], Shishir Kumar[a], Steven Lawrence Fernandes[b]

[a] Department of Computer Science & Engineering, Jaypee University of Engineering & Guna (M.P.) , India
[b] Department of Electronics & Communication Engineering, Sahyadri College of Engineering & Management, Mangalore, Karnataka, India

ARTICLE INFO

Article history:
Received 16 July 2016
Received in revised form 5 December 2016
Accepted 28 January 2017
Available online 1 February 2017

Keywords:
Emotion extraction
Machine learning
Text mining
Twitter
Classification
Natural language processing

ABSTRACT

Extraction of Emotions from Multilingual Text posted on social media by different categories of users is one of the crucial tasks in the field of opining mining and sentiment analysis. Every major event in the world has an online presence and social media. Users use social media platforms to express their sentiments and opinions towards it. In this paper, an advanced framework for detection of emotions of users in Multilanguage text data using emotion theories has been presented, which deals with linguistics and psychology. The emotion extraction system is developed based on multiple features groups for the better understanding of emotion lexicons. Empirical studies of three real-time events in domains like a Political election, healthcare, and sports are performed using proposed framework. The technique used for dynamic keywords collection is based on RSS (Rich Site Summary) feeds of headlines of news articles and trending hashtags from Twitter. An intelligent data collection model has been developed using dynamic keywords. Every word of emotion contained in a tweet is important in decision making and hence to retain the importance of multilingual emotional words, effective pre-processing technique has been used. Naive Bayes algorithm and Support Vector Machine (SVM) are used for fine-grained emotions classification of tweets. Experiments conducted on collected data sets, show that the proposed method performs better in comparison to corpus-driven approach which assign affective orientation or scores to words. The proposed emotion extraction framework performs better on the collected dataset by combining feature sets consisting of words from publicly available lexical resources. Furthermore, the presented work for extraction of emotion from tweets performs better in comparisons of other popular sentiment analysis techniques which are dependent of specific existing affect lexicons.

1. Introduction

Emotion expression plays a vital role in various part of everyday communication. In past, various measures have been used to evaluate it, through a combination of indications such as facial expressions, gestures, and actions etc. Emotions extraction using facial, gestures and action are the part of digital image processing and computer vision [1]. Emotions extraction is more difficult from texts especially from multi-languages texts, like in posts on social media and customers' reviews. This type of data has presence of ambiguity and complexity of words in terms of meaning make them more difficult. Factors such as users writing style, politeness, irony, variability in language is one of the important problems in extraction of emotions [2]. A wide variety of state-of-art work has been carried out in the domain of opinions mining and sentiment analysis but limited research are focused on detection/extraction of emotions in multi-language text.

In English vocabulary, some words express emotion explicitly, whereas other words can be used to get across emotion implicitly depending on the context [3]. Emotion detection in the text has recently attracted the scientific community to explore meaningful inferences hidden in the data and help in decision-making [4]. Many authors classify emotions in multiple classes for a better understanding, like Strapparava and Valitutti have classified emotional words into two classes’direct affective words’ and’indirect affective-words [2]. Emotion research is important for building affective interfaces. These affective interfaces provide better user experience in following areas such as Human–Computer Interaction (HCI), Text-to-Speech (TTS) synthesis systems and Computer-Mediated Communication (CMC) [5]. Computational techniques related to emotion extraction present in social media

[*] Corresponding author.
E-mail addresses: vinay2588@gmail.com (V.K. Jain), dr.shishir@yahoo.com (S. Kumar), steven.ec@sahyadri.edu.in (S.L. Fernandes).

http://dx.doi.org/10.1016/j.jocs.2017.01.010
1877-7503/© 2017 Elsevier B.V. All rights reserved.

Table 1
Basic emotion categories recognized by the different researchers.

Authors	Emotion class	Emotion labels
Tomkins [9]	8	Joy, anguish, fear, surprise disgust, interest, shame, anger
Izard [10]	10	Enjoyment, shame,fear, anger,surprise, interest, sadness,shyness, guilt, disgust
Plutchik [11]	8	Joy, sorrow, anger, fear, disgust, surprise, acceptance, anticipation
Ortony et al.et al. [12]	6	Joy, surprise fear, anger,sadness, disgust
Ekman [8]	6	Happiness, sadness, anger, disgust, surprise, fear
Muni [13]	8	Jugupsa (Disgust), Hasya (Mirth), Krodha (Anger), Rati (Love), Utah (Energy), Bhaya (Terror), Vismaya (Astonishment), Soka (Sorrow)

have paying attention on basis of multiple emotion modalities [6]. However, only limited work has been done in developing automatic emotion recognition system [4,6].

The multilanguage text contains emotional words of different languages and extraction of these emotional words definitely improve emotion identification ratio [7]. In most of the available literature, theses words are treated as stop words in social media data [7]. This paper presented an advanced framework for automatic detection of emotions in Multilanguage text data. The emotion models used for development of proposed framework deals with linguistics and psychology. Proposed framework uses Machine Learning techniques for learning and validation and effective pre-processing Natural Language Processing (NLP) techniques for better extraction of emotions existing in the text.

This paper uses the concept of emotion model given by Ekman [8] as a basis with multiple feature sets to deal with multilingual data. The text under study comprises data collected from Twitter in three different domains such as Political election, Healthcare, and Sports. The first task is to collect real-time data consisting of relevant keywords. Through this paper, a novel technique based on RSS (Rich Site Summary) feeds to collect keyword which has been used for real-time data collection of events, has been introduced. Tweets containing images and emoticons are not considered under the scope of proposed approach. The effective pre-processing technique has been used to filter out irrelevant words and preserving words representing emotion of other languages. The classification of the dataset has been performed using popular machine learning techniques.This work represents the first systematic evaluation of emotion detection in real-time multilingual data in multiple domains. Another key contribution of the presented work is the practical application of emotion models in comparison of corpus-driven approach which assigns affective orientation or scores to words and word frequencies.

The rest of the paper has been organized as follows. State-of-art methods have been presented in Section 2. Proposed data collection methodology has been presented in Section 3. The problem formulation, existing methods, and proposed framework of emotion extraction system have been presented in Section 4. Experimental setup and outcomes with discussions have been presented in Sections 5 and 6. In Section 7 advantages of proposed approach over state of art, methods have been identified. Finally, precise conclusions and scope of future work are mentioned in Section 7.

2. Related work

Nowadays, a lot of research articles have been published for analyzing sentiments in social media data in multiple domains. This literature review section discussed emotion extraction methods and sentiment classification methods related to different domains like election prediction, healthcare, and sports analytics.

2.1. Emotion extraction methods

Researchers have investigated basic human emotions in different categories that are accepted universally. A number of related works in the field of emotion identification, reported in the literature, has been presented in Table 1.

Ekman's emotion theory [8] is the most popular and widely used approaches related to emotion recognition. Human emotions can be recognized using, speech, facial expression, gestures, and writings [4,5]. Research in emotion identification has focused on all these aspects [14]. Finding accurate emotions in the text contains evident in the vast body of research work related to different fields of psychology, social sciences, linguistics, Human–Computer Interaction (HCI) and communication. Table 2 outlines the significance of emotion detection and recognition techniques used in multiple domains.

2.2. Sentiment classification methods

The field of sentiment analysis recently witnessed a large amount of interest from the scientific community [37–40]. It deals for automatically determining the polarity of a textual data based on polarity, whether it is positive, negative, or neutral. More recently, much effort has been invested into the development of sentiment analysis methods in comparison to emotion extraction across multiple domains like movie and product reviews, election result prediction; disease outbreak, sports, stock market etc. [36,41,42]. In this paper, sentiment analysis techniques used in three specific domains such as Political election, Healthcare and Sports has been presented.

Tracking public sentiment during elections is a hot research area and prediction based on these sentiments is effective in comparison to survey-based methods. Nowadays, every election campaign has an online presence and users use social media platforms to express their sentiments and opinions towards political parties, leaders, and important topics during election [43]. Optimistic results regarding the predictive capacity of social media towards the election results in geographical regions are illustrated in Table 3. According to most of the authors, better prediction depends on the quality of data and merits of data collection methodologies [44]. If the data collected is not much relevant towards the event then the outcomes may be inappropriate [45]. Most of the authors focused on corpus-based feature and sentiment orientation technique for predicting the election outcomes, but emotion detection of social media users has not been taken into suitable consideration.

Another popular domain in the field of sentiment analysis is healthcare. Corpus-based features and sentiment based techniques have been providing a rich source of information for detecting and forecasting disease outbreak in all around the world [42]. Chew [71] used specific keywords related to outbreak detection in 2009 H1N1 pandemic.Hu et al. [72] used web services provided by Google related to influenza epidemic using specific keywords. Lampos and Cristianin [73] used content based methods with statistical methods to monitor and measure public perceptions. They also analyzed levels H1N1 pandemic. Chunara et al. [74] detected cholera outbreak using Twitter. Aramaki et al. used Support Vector Machine for predicting influenza rates in Japan [75].Stewart and Diaz [76] developed a real-time data analysis of disease using social media with an early warning system.Bodnar et al. [77] applied various

Table 2
Emotion detection and recognition techniques.

Authors	Techniques	Explanation
Osgood et al. [15]	Factor analysis of texts	Factor for evaluation is (good or bad), Factor for activity (passive or active) and Factor for potency (weak or strong)
Jakobson [16]	Emotive function	Creative use of words for communicating emotion
Watson and Tellegen [17]	Scaling positive and negative affect from low to high	Pleasantness–unpleasantness and engagement–disengagement
Johnson-Laird and Oatley [18]	Basic emotions words	Given 590 English emotion words
Fellbaum [19]	WordNet	Automatically acquire emotion-related words for experiments
Bradley and Lang [20]	Factor scores	Presents happiness factors
Kamps and Marx [21]	Relative weight of words	Weighted emotion words
Liu et al. [22]	Mapped the affective content	Represent the document using color bar
Martin and White [23]	Appraisal framework	Conveying attitudes, judgments, and emotions
Alm et al. [24]	Utilizing all senses of all words in the synsets that contain the emotion adjectives	Retrieve similar words from the WordNet
Mishne [25]	Classification of text	Classify LiveJournal blogposts
Strapparava and Valitutti [2]	Affective extension of WordNet	Automatically detect emotion in text
Strapparava et al. [26]	Divided into two categories	Direct affective words (explicit) and indirect affective words (implicit) categories
Mihalcea and Liu [27]	Divided into two categories	Happiness and sadness
Zhang et al. [28]	Affect in the characters' speeches	Classify affect words
Mihalcea and Strapparava [5]	Automatic recognition of humor in texts	Humor detection in text
Read [29]	Appraisal annotation of a corpus	Used book reviews
Neviarouskaya et al. [30]	Augmenting online conversations	Graphical representation of text
Ghazi et al. [5]	Hierarchical classification	Based on six Ekman [8] emotions
Chaumartin [31]	Linguistic rule-based	Developed UPAR7 System
Dung et al. [32]	Human mental states	Hidden Markov Model (HMM)
Lily Dey et al.et al. [33]	Extracting emotions from real time	Chat messenger
Shadi Shaheen et al. [34]	Intermediate emotional data representation	Based on syntactic and semantic structure
Jordon [35]	Geospatial analysis of Twitter sentiment	Geospatial similarly analysis in political elections
Aman and Szpakowicz [36]	Corpus-based unigram features with emotion-related features	Lexicon is automatically built using Roget's Thesaurus and words extracted from WordNet-Affect

Table 3
Contribution of different authors for election prediction.

Authors	Country	Election type	Method(s)
O'Connor et al. [46]	US	Presidential	Sentiment analysis using word frequencies
Tumasjan et al. [47]	Germany	Federal	Corpus-based/Hashtags
Hopkins and King [48]	US, France, Italy	Presidential	Sentiment classification based on lexical induction
Tumasjan et al. [49]	Germany	Federal	Sentiment analysis of tweets
Diakopoulos and Shamma [50]	US	Presidential	Demonstrate visuals and metrics of tweets
Choy et al. [51]	Singapore	Presidential	Reweighting techniques, corpus-based & sentiment analysis
Bermingham and Smeaton [52]	Ireland	General	Corpus-based & sentiment analysis
Metaxas [53]	US	General	Multiple methods
Conover et al. [54]	US	US midterm	SVM
Maynard and Funk [55]	UK	Pre-election	NLP techniques
Larsson and Hallvard [56]	Sweden	General	Identifying user types
Choy [57]	US	Presidential	Corpus-based
Skoric et al. [58]	Singapore	General	Corpus-based
Tjong et al. [59]	Dutch	Senate	Sentiment analysis
Johnson [60]	US	Presidential	Corpus-based
Marquez et al. [61]	US	Presidential	Time series of Twitter messages
Shi et al. [62]	US	Republican Presidential	Corpus-based and sentiment analysis
Soler et al. [63]	Spain	Regional and general	Corpus-based
Contractor and Faruquie [64]	US	Presidential	NLP techniques
Nooralahzadeh [65]	US & French	Presidential	Sentiment analysis
Kim et al. [66]	US	Presidential election	Iterative topic modeling algorithm
Bakliwal et al. [67]	Ireland	General	3-Class sentiment classification
Then and Ghanem [68]	UK	Members of Parliament	Automatic identification
Song et al. [69]	Korea	Presidential	Text-mining techniques
Vaccari et al. [70]	Italian	General	Corpus-based

classification techniques for detecting influenza. Parket et al. [78] developed a framework to tracks the levels using trends extracted from Twitter. Jain and Kumar [7] have tracked levels of Influenza-A (H1N1) during 2015.

Every Sports event evokes strong emotions of the public which help in monitoring public sentiment towards the team or the players. Due to rise in social media data towards sports, sports-related companies have interested in mining data and extracted meaningful inference for improving their marketing strategy for products and services. For example, many fans seeking information related to particular team or player and show their interests using by liking or commenting will become the potential buyer for

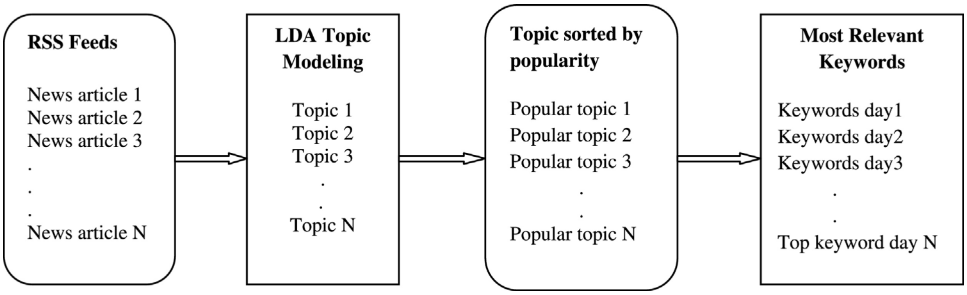

Fig. 1. Selection procedure for topics and keywords from RSS feeds.

product and services. Recent research showed that maximum users in the age group of 18–35 years follow players and teams in social media platforms and regularly posts, share or link related to sports contents [79].

A number of methods have been used in past to forecast success (winners and losers), both for single games and team games in the sport. These events have a great uncertainty towards the results if teams and players are strong contenders. Limited research is carried out to forecast results using social media. According to Wang, a supporter of the team and players use social media to express their emotion and opinions [80] .Yu and Wang [81] use twitter data for FIFA world cup 2014 to analysis the emotions of users and also describe event-based tweets responses. Sentiment analysis techniques have been used for prediction of outcomes of English Premier League (EPL) with promising results [82,83]. Sinha et al. [84] used *n*-grams from Twitter data sets to predict outcomes of the National Football League (NFL) and compared it with other simple statistics methods. Lock and Nettleton [85] applied machine learning technique to classify tweets and also used situational variables. UzZaman et al. [86] used a framework (TwitterPaul) to extract tweets and find the outcome of FIFA World cup 2015.Harnessing the emotions of the crowd, from the social media data for making predictions towards sports event need more effort. This paper suggests a new dimension for sports event analytics using emotions posted in social media data.

3. Proposed data collection methodology

In this section, an intelligent technique for data collection has been presented. The important variable for data collection from social media data are keywords, which helps in identification of relevant tweets. Most research for keyword selection is based on popular terms corresponding to the event [46,49,52,57,58,62]. Methodology for data collection is different from other author's techniques; here only those keywords which are trending and dynamic are considered.

The process of keywords collection starts with the extraction of newspapers RSS feeds. The topic modeling technique, Latent Dirichlet Allocation (LDA) given by Blei [87], has been used to filter out most occurred topics and keywords. These extracted keywords with trending hashtags are used for data collection using Twitter API. The methodology used for data collection is presented in Fig. 1.

According to Blei [87] model, the selected topics *k* has to be fixed a priori. For a document $w = (w_1,w_2. . ., w_N)$ of a corpus *D* (collection of RSS feeds) having the presence of *N* words from a vocabulary having *S* different terms, $w_i \in \{1,. . ., S\}$ for all $i = 1,. . ., N$.

Every word contained in a tweet is important in decision making, so pre-processing of these tweets is an important task because these messages are full of slang, misspellings and words from other languages. In order to tackle the problems with the noise in texts, normalization of tweets is performed by intelligent text pre-processing steps like tokenization, stop-words removal, stemming, lemmatization, feature weighting, dimensionality reduction, frequency based methods proposed by Bao et al. [88].

For testing the effectiveness of the proposed emotion detection system, a large-scale dataset is used in the experiment. For election dataset, data collection process started during assembly election of Delhi (India). The dataset contains 1085721 tweets after filtering of 390425 re-tweets or duplicates, 703521 tweets have remained. This dataset is divided into three parts based on different time periods and is presented in Table 4.

For healthcare dataset, data has been collected during an epidemic of Influenza-A H1N1 in 2015.The data set consists of 91495 tweets. Keywords used for data collection are #SwineFlu, #, H1N1, #h1n1,#Flu,# influenza, #Fluvirus.

For sports event dataset, data has been collected for Indian Premier League (IPL) 2015 Cricket tournament, which is one of the famous cricket tournaments in India. It is an annual cricketing competition that is held in India from April to May every year between eight teams. This event is one of the widely watched events in India during 2015 and a lot of zest surrounds this contest. Keywords used for data collection are: #IPL, #IPL2015, #IndiaKaTyohaarBegins,#IPL8, #PepsiIPL2015, #PepsiIPL. Every cricket match during IPL 2015 season has corresponding hashtags such as #KKRvMI, #CSKvMI, #RRvKKR, #KKRvMI, #CSKvDD, #KXIPvRR, #CSKvRR etc. and is used for data collection.

4. Proposed methodology

In this section, proposed Emotion extraction framework, emotion models with annotation of general terms and feature groups used in the framework has been presented.

4.1. Proposed emotion extraction framework

Emotion extraction in the text is considered as a classification problem. Emotion labels have been assigned to a tweet from a group of multiple emotion labels. Proposed framework for emotion extraction in tweets represents as:

Let *t* is a tweet and *k* is an emotion label. Considering $e = \{e_1, e_2, e_3. . . . . e_n\}$ is a set of *n* possible emotion categories. The main aim is to label *t* 'tweet' with best emotion label *k* from the set of multiple emotion labels, where $k \in \{e_1, e_2, e_3 . . . . . . e_n, \text{neutral}\}$.

Classification of data sets has been performed in two steps. Firstly, the dataset is divided into two basic classes, namely, emotion and non-emotion using Support Vector Machine (SVM) and Naive Bayes (NB).Secondly, effective fine-grained emotion classification has been performed using Support Vector Machine (SVM) and Naive Bayes (NB) and is presented in Fig. 2.

Emotion labeling is reliable if there is more than one judgment for each label [4,36,41]. Emotion labeling has been done by four judges who manually annotated the corpus based on Aman and Szpakowicz [4] method. In this process, each tweet affinity was subject to two judgments. The annotators also identify spans of

Table 4
Tweets collected during different time intervals with top hashtags.

Time period	No. of tweets	Top keywords
20-Nov-2014 to 20-Dec-2014	166380	#delhielections, #delhipolls, #Kejriwal #Arvind,#CrazyKejaria,#aap, #kejriwal, #bhagodakejri
20-Dec-2014 to 20-Jan-2014	212448	#CleanPolitics, #vote4aap,#BJPdelhi,,#missiondelhi #Mufflerman #*DelhiVotes*#Kejriwal4Delhi,#AAPKiDilli #MufflerMan
20-Jan-2015 to 06-Feb-2015	312694	#5saalkejriwal#abkibaarbedisarkaar vs #IronLady, #Kiran bedi,#kejriwal4delhi, #congresskidilli, #aapkamanifesto

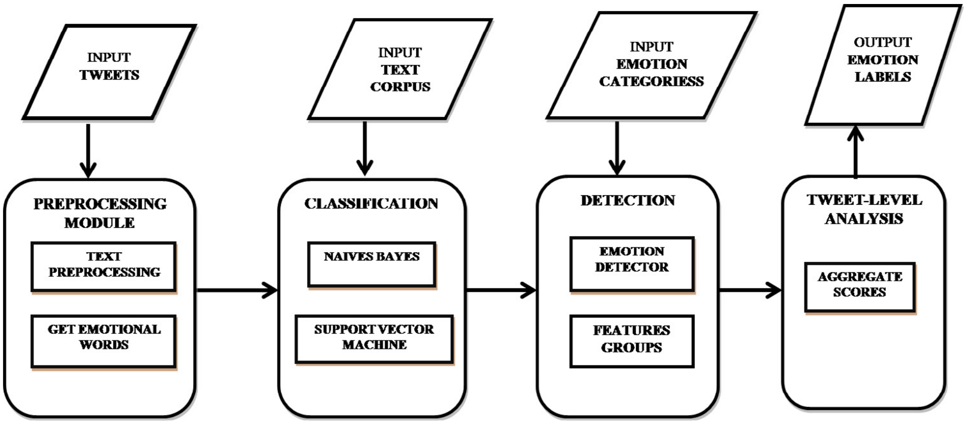

Fig. 2. Proposed framework for emotion extraction.

text (individual words or strings of consecutive words) that convey emotional content in a tweet. The annotators received no training, though they were given samples of annotated tweets to illustrate the kind of annotations required. This process of annotation gives a broad range of affect-bearing lexical tokens and syntactic phrases. The annotated data was prepared over a period of 6 months.

The annotators were required to label each tweet with the appropriate emotion category, which describes its affective content. In addition to Ekman emotion category two new categories (mixed emotion and no-emotion) has been formed for a better understanding of data. Ekman [8] emotion model is used for fine-grained emotion classification and explained in Section 4.2. Multiple feature groups are used for classification of data set in which has been presented in Section 4.3.

The interpretation of the presence of emotion in the text is highly subjective, which leads to the disparity in the annotations by different judges [41]. The difference in skills, technique and focus of the judges, and ambiguity in the annotation guidelines and in the annotation task itself also contribute to a disagreement between the judges [4,41]. We seek to find how much the judges agree in assigning a particular annotation by using metrics that quantify these agreements. Cohen's kappa which is a popular technique used to compare the extent of consensus among judges in classifying items into known mutually exclusive categories. The average value calculated for kappa in the case of Pair-wise agreement in emotion/non-emotion labeling is 0.75. Table 5 shows the value of kappa for each of these emotion categories for each annotator pair.

For evaluation purposes, a baseline system has been developed. The system that counts the emotion words of every category in a tweet. The category with the largest number of emotional words to finding in a tweet has been assigned to it. For obtaining prior knowledge about emotion-bearing words, seeds words related to emotions words have been extracted from WordNet-Affect [2], Hindi WordNet-Affect (HWNA) [89] and SentiwordNet [90] into basic emotion categories. Ekman [8] model is used for classification of emotion present in tweets. Similar classification of emotions categories such as surprise, happiness, sadness, disgust, anger and fear which are used by Aman [41].

Table 5
Emotion categories used in annotation and pairwise agreement in emotion categories.

Emotion category	Label	Pair-wise agreement in emotion categories
Happiness	hp	0.76
Sadness	sd	0.65
Anger	ag	0.67
Disgust	dg	0.66
Surprise	sp	0.59
Fear	fr	0.77
Mixed emotion	me	0.42
No emotion	ne	0.74

The annotation scheme used for labeling the tweets is represented in Table 5. Each tweet should be labeled with the appropriate emotion category, which best described its affective content. No-emotion category was added to for those tweets that do not have any emotion content in them.

4.2. *Defining feature sets*

For defining feature sets for automatic classification of emotions in tweets, it is important to consider emotional words which distinctly characterize emotions in multilingual text. Three publicly available lexical resources such as WorldNet-Affect (WNA) [2], Hindi WordNet-Affect (HWNA) [89] and Senti-WorldNet [90] are used for creating features sets. Three feature sets are formed which contains appropriate features that help in distinguishing multi-language emotional words from non-emotional words. Following are the four feature groups used in experiments:

(i) Group 1: Features of WorldNet-Affect (WNA) [2]
(ii) Group 2: Features of WorldNet-Affect (WNA) [2] and Senti-world Net [90]
(iii) Group 3:Features of WorldNet Affect with Hindi WordNet-Affect (HWNA) [89]

(iv) Group 4:Combined features of WNA [2]+HWNA [89]+Sentiword Net [90]

5. Experiments

In this section, performance analysis of the proposed system for emotion extraction with corpus-based features has been evaluated on collected datasets. Firstly, Corpus-based feature analysis present in the datasets has been exploited. Secondly, experimental analysis of proposed emotion extraction framework using multiple datasets has been evaluated.

5.1. Corpus-based feature analysis

In this section, Corpus-based features exploit the statistical features based on the distribution of *n*-grams present in the datasets. In the experiments, unigrams ($n = 1$) are used as features. Important key terms like Political parties names, Chief Minister (CM), Healthcare terms, Team names, Player's names etc. are used as unigrams are extracted using count based technique and presented in Tables 6–9.

In the case of healthcare dataset of Influenza-A (H1N1), the following are the most occurred unigrams using count- based technique has been presented in Table 8.

In the case of Sports data collected during IPL 2015 cricket tournament, considering players and team names used as unigrams and are filtered out using count based technique and presented in Table 9.

Table 6
Total tweets count for political party.

Total tweets counts: 420705			
Political party	Party 1 (AAP)	Party 2 (BJP)	Party 3 (Congress)
Total	211500	156188	53017

Table 7
Total tweets count for CM candidate.

Total tweets counts: 216164			
CM candidate	Candidate 1 (Arvind Kejriwal)	Candidate 2 (Kiran Bedi)	Candidate 3 (Ajay Maken)
Total	168183	44426	3555

Table 8
Top 20 most frequently occurring unigrams.

Unigram	Counts	Unigram	Counts
SwineFlu	12015	Sick	1063
Flu	7471	Health	1053
H1N1	6337	Mumbai	804
Swine	5318	Outbreak	692
Flu	3365	Influenza	671
Swine	3294	Rajasthan	624
Virus	2299	Kashmir	604
Deaths	1975	Feel	553
Swine flu	1788	Andra	425
Gujarat	1117	Maharashtra	388

Table 9
Tweets counts of popular players and teams.

Top players	Counts	Cricket team	Counts
Player 1 (V Kohli)	19823	Team 1 (KKR)	56174
Player 2 (RG Sharma)	15529	Team2 (CSK)	52340
Player 3 (AB de Villiers)	13111	Team 3 (MI)	47694
Player 4 (MS Dhoni)	10934	Team 4 (DD)	42177
Player 5 (CH Gayle)	9054	Team 5 (RCB)	37456

Table 10
Results of ten-fold cross validation for emotion/non-emotion classification.

Features group	Dataset 1 (Election)		Dataset 2 (Healthcare)		Dataset 3 (Sports)	
	Naives	*SVM*	*Naives*	*SVM*	*Naives*	*SVM*
Group 1	71.65%	71.33%	68.45%	70.33%	69.45%	71.23%
Group 2	72.16%	70.58%	71.16%	70.58%	70.16%	70.18%
Group 3	72.70%	73.89%	72.12%	73.89%	71.70%	72.89%
Group 4	74.80%	75.61%	73.90%	74.10%.	72.89%	73.76%

5.2. Emotion extraction framework

In this section, two experiments have been performed for classifying tweets. Firstly, classification of the dataset into two basic categories, namely, emotion (EM) and non-emotion (NE) has been performed. Secondly, proposed an effective technique for fine-grained emotion classification using Emotion model given by Ekman [8]. For the classification experiments, Naive Bayes (NB) and Support Vector Machines (SVM) has been used. Four different sets of experiments were performed to test the effectiveness and contribution of the different feature groups given in Section 4.2.

A comparative performance evaluation of Naïve Bayes (NB) and Support Vector Machine (SVM) in terms of correctly classify emotions containing tweets has been examined. The results are explained in terms of precision, recall, accuracy, and *F*-measure and represented in Table 10. For building the training and testing sets, Ten-fold cross-validation experiments have been conducted using the NB and SVM. The data set has been randomly split into *k* "folds". For each '*k*' folds in the dataset, build the model on '$k-1$' folds of the data set. Then, test the model to check the effectiveness for '*k*th' fold. Record the error of each of the predictions. Repeat this until each of the '*k*' folds has served as the test set. The average of '*k*' recorded errors is called the cross-validation error and will serve as the performance metric for the model. For each fold we calculated tp, tn, fp, and fn. Then we calculated the accuracy, precision, recall and *F*-score for each test .Precision and recall have been calculated using Multi-class classification of text techniques which are presented in Refs. [4,6,14,24,30,31,33,41]. The *F*-measure provides the overall performance of a classifier and is calculated using the following formula given by Eq. (1).

$$F - \text{measure} = \frac{2(\text{Precision})(\text{Recall})}{\text{Precision} + \text{Recall}} \quad (1)$$

On Comparing, the performance of the SVM classifier was found better than the Naïve Bayes classifier in maximum test run cases. The highest accuracy achieved was 72.81%. The combination of all feature groups provides a better result with an indication that the combination of features is needed to improve emotion classification results in multilingual datasets.

To evaluate the usefulness of presented framework, quantitative analysis has been performed by comparing the performance based on feature groups described earlier in Section 4.3. NB and SVM method has been used for predicting the emotion category of the tweets in the dataset. Classifiers with features of each emotion class have been trained using seed words taken from multiple languages. In all experiments, a tweet was represented by a vector having values which indicate the occurrence of each feature. Tables 11 and 12 present results of ten-fold cross-validation experiments conducted using the Naïve Bayes and SVM implementation.

Four experiments have been performed using feature groups. The performance using the SVM classifier was found better than Naïve Bayes classifier in maximum cases in experimented data sets. Text classification deals with high dimensionality of feature space because of it various learning algorithms do not work with high dimensional feature space. Tables 11 and 12 clearly showed that SVM outperform in maximum cases in experimented data set.

Table 11
Results of fine-grained classification using Naive Bayes.

Features group	Class	Dataset 1 (Election)			Dataset 2 (Healthcare)			Dataset 3 (Sports)		
		Precision	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*
Group 1	Happiness	0.713	0.367	0.485	0.649	0.297	0.408	0.716	0.325	0.447
	Sadness	0.426	0.349	0.384	0.401	0.372	0.386	0.316	0.341	0.328
	Anger	0.314	0.312	0.313	0.314	0.332	0.323	0.279	0.298	0.288
	Disgust	0.539	0.340	0.417	0.620	0.309	0.412	0.579	0.290	0.386
	Surprise	0.317	0.240	0.273	0.337	0.243	0.282	0.477	0.294	0.364
	Fear	0.528	0.379	0.441	0.538	0.374	0.441	0.628	0.352	0.451
	No-emotion	0.414	0.062	0.108	0.312	0.082	0.130	0.504	0.013	0.025
Group 2	Happiness	0.677	0.429	0.525	0.587	0.219	0.319	0.597	0.308	0.406
	Sadness	0.348	0.319	0.333	0.378	0.349	0.363	0.438	0.217	0.290
	Anger	0.420	0.181	0.253	0.421	0.281	0.337	0.491	0.239	0.322
	Disgust	0.614	0.179	0.277	0.594	0.189	0.287	0.692	0.267	0.385
	Surprise	0.368	0.256	0.302	0.318	0.316	0.317	0.354	0.237	0.284
	Fear	0.682	0.411	0.513	0.689	0.401	0.507	0.654	0.383	0.483
	No-emotion	0.297	0.026	0.048	0.257	0.022	0.041	0.298	0.024	0.044
Group 3	Happiness	0.681	0.376	0.484	0.710	0.342	0.462	0.580	0.379	0.458
	Sadness	0.362	0.334	0.347	0.312	0.394	0.348	0.378	0.447	0.410
	Anger	0.250	0.376	0.300	0.266	0.336	0.297	0.390	0.339	0.363
	Disgust	0.317	0.298	0.307	0.317	0.352	0.334	0.477	0.319	0.382
	Surprise	0.246	0.317	0.277	0.251	0.271	0.261	0.386	0.265	0.314
	Fear	0.340	0.466	0.393	0.359	0.396	0.377	0.450	0.446	0.448
	No-emotion	0.463	0.085	0.144	0.451	0.095	0.157	0.391	0.066	0.113
Group 4	Happiness	0.658	0.394	0.493	0.538	0.404	0.461	0.448	0.377	0.409
	Sadness	0.341	0.522	0.413	0.391	0.512	0.443	0.451	0.433	0.442
	Anger	0.251	0.308	0.277	0.278	0.328	0.301	0.378	0.355	0.366
	Disgust	0.423	0.348	0.382	0.502	0.418	0.456	0.522	0.378	0.438
	Surprise	0.265	0.316	0.288	0.303	0.216	0.252	0.393	0.288	0.332
	Fear	0.352	0.386	0.368	0.316	0.396	0.352	0.466	0.446	0.456
	No-emotion	0.470	0.092	0.154	0.503	0.079	0.137	0.482	0.083	0.142

Table 12
Results of fine-grained classification using SVM.

Features group	Class	Dataset 1 (Election)			Dataset 2 (Healthcare)			Dataset 3 (Sports)		
		Precision	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*
Group 1	Happiness	0.750	0.689	0.718	0.762	0.664	0.710	0.819	0.825	0.822
	Sadness	0.589	0.325	0.419	0.632	0.392	0.484	0.648	0.411	0.503
	Anger	0.784	0.367	0.500	0.654	0.382	0.482	0.647	0.428	0.515
	Disgust	0.872	0.493	0.630	0.765	0.472	0.584	0.754	0.513	0.611
	Surprise	0.583	0.367	0.450	0.876	0.382	0.532	0.837	0.449	0.584
	Fear	0.989	0.465	0.633	0.832	0.472	0.602	0.877	0.557	0.681
	No-emotion	0.631	0.352	0.452	0.532	0.356	0.427	0.593	0.472	0.526
Group 2	Happiness	0.552	0.782	0.647	0.765	0.574	0.656	0.784	0.772	0.778
	Sadness	0.684	0.555	0.613	0.578	0.345	0.432	0.589	0.335	0.427
	Anger	0.758	0.766	0.762	0.634	0.456	0.530	0.659	0.376	0.479
	Disgust	0.839	0.837	0.838	0.776	0.345	0.478	0.749	0.497	0.598
	Surprise	0.668	0.573	0.617	0.743	0.288	0.415	0.798	0.523	0.632
	Fear	0.877	0.840	0.858	0.898	0.468	0.615	0.839	0.810	0.824
	No-emotion	0.928	0.348	0.506	0.443	0.227	0.300	0.458	0.598	0.519
Group 3	Happiness	0.649	0.885	0.749	0.864	0.785	0.823	0.866	0.695	0.771
	Sadness	0.827	0.660	0.734	0.522	0.365	0.430	0.555	0.670	0.607
	Anger	0.946	0.559	0.703	0.665	0.470	0.551	0.688	0.444	0.540
	Disgust	0.846	0.661	0.742	0.647	0.484	0.554	0.699	0.488	0.575
	Surprise	0.697	0.294	0.414	0.784	0.393	0.524	0.722	0.355	0.476
	Fear	0.491	0.603	0.541	0.876	0.573	0.693	0.833	0.522	0.642
	No-emotion	0.706	0.702	0.704	0.554	0.526	0.540	0.569	0.544	0.556
Group 4	Happiness	0.993	0.598	0.746	0.821	0.677	0.742	0.844	0.687	0.757
	Sadness	0.895	0.696	0.783	0.755	0.455	0.568	0.677	0.444	0.536
	Anger	0.860	0.586	0.697	0.730	0.475	0.576	0.699	0.477	0.567
	Disgust	0.542	0.738	0.625	0.832	0.434	0.570	0.655	0.495	0.564
	Surprise	0.823	0.699	0.756	0.833	0.463	0.595	0.764	0.487	0.595
	Fear	0.788	0.873	0.828	0.638	0.578	0.607	0.877	0.545	0.672
	No-emotion	0.697	0.795	0.743	0.697	0.698	0.697	0.592	0.679	0.633

6. Results

In this section, performance analysis of the proposed emotion extraction system has been evaluated on collected datasets. The important meaningful inference drawn from datasets has been presented. Different test data sets are used for predicting results on the basis of events.

In the case of election outcome prediction, two test cases based on party name and candidate name has been formed. In the first case, emotion extraction model has been applied to derive the emotion towards CM candidate. Similarly, in the second case, political party has been taken. A randomly collected dataset of 20,000 tweets has been used for the experimental study.

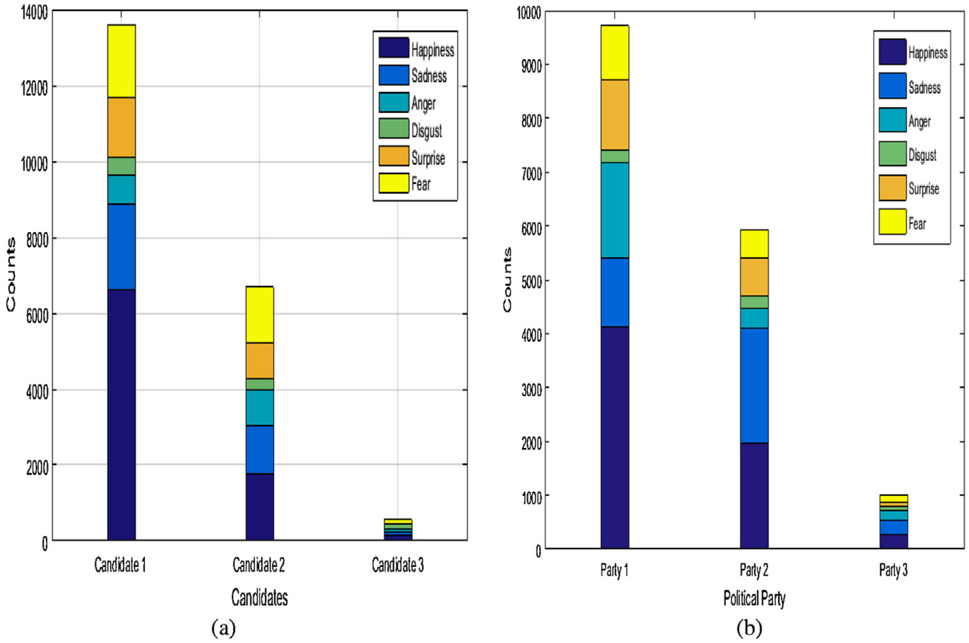

Fig. 3. Results of proposed emotion extraction method. (a) Case of CM candidates; (b) Case of political party.

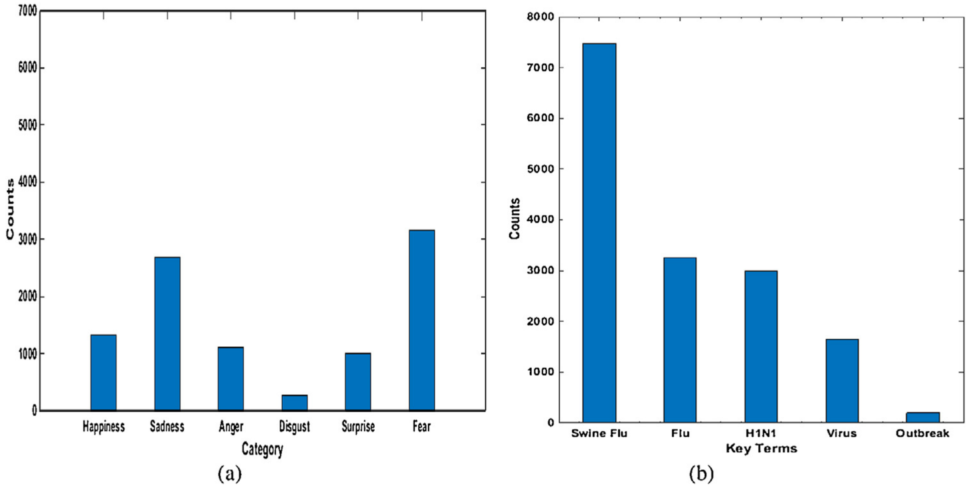

Fig. 4. Results of emotion extraction method with important word frequencies. (a) Emotions related to Influenza-A (H1N1); (b) Key terms frequencies related to Influenza-A (H1N1).

For election outcome prediction, public emotion related to candidates and political parties has been presented in Fig. 3. On the basis of results; it can be observed that public opinion is in favor of Candidate 1 and Party 1 in comparison to other competitors. In the case of Healthcare, analysis of emotion towards Influenza-A (H1N1) and word frequencies of important key terms has been examined in Fig. 4.

In the case of social media based healthcare data analysis, authors used key terms frequencies based analysis for detection of outbreaks [71–75]. Detecting emotions during epidemics or healthcare event based on emotion categories present variability in public emotions which lead to help in decision making. For example, Fig. 4(a) gives a clear indication of public emotions during Influenza-A (H1N1) 2015 in India. Emotion-based analysis will help government and health agencies in decision making as compared to key-term based frequency analysis presented in Fig. 4(b).

In the case of Sports analytics, due to rising in social media data towards sports, sports-related companies are interested in mining this data and extracted meaningful inference from them which help in improving their marketing strategy for products and services. For example, team franchises paying a huge amount of money to players in addition to their travel and maintenance costs during the tournament. The franchise can improve their revenues by ticket sales, advertisements, or by selling merchandise. Public emotion related to teams and team players will help them in decision making. In concern of business analytics in the sports, some important insight of public emotion related to teams or players are examined and presented in Fig. 5(a) and (b). It can be easily observed that public emotions are in favor of specific players and teams from Fig. 5.

Experiments presented in this section clearly showed that emotion-related features give a better understanding of social media data and will be applicable in other domains.

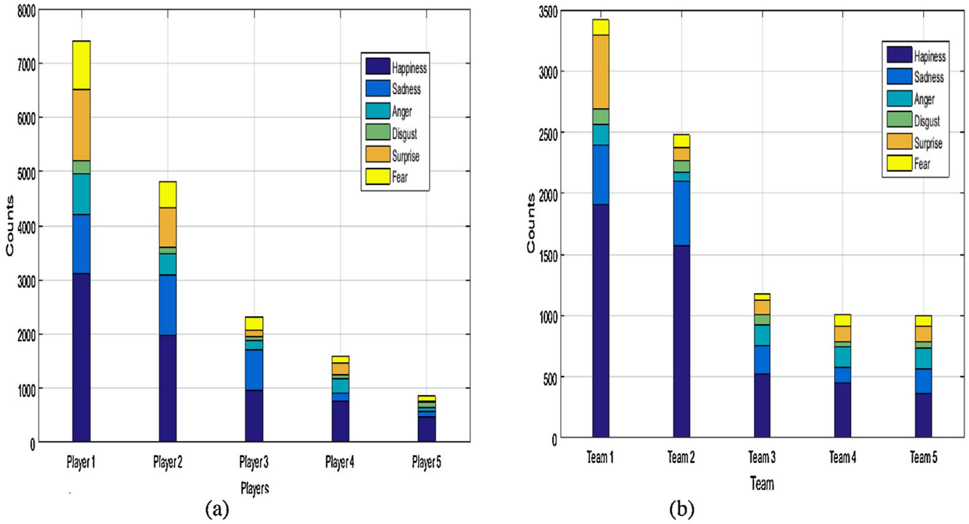

Fig. 5. Results of proposed emotion extraction method. (a) Case of team players (b) Case of teams.

7. Advantages of proposed work

The proposed models have been used in multiple data-driven applications which focused on the hidden information contained in the text. An application such as topic-based text categorization, summarization, question answering systems, and information retrieval systems can be improved using proposed method.

Emotion research is widely used in developing affective interfaces which provide appropriate emotional responses and facilitate online communication through animated affective agents [91,92]. These interfaces are suitable for improvisation of user experience in Computer-Mediated Communication and Human–Computer Interaction.

Extraction of emotions from social media towards events related to politicians, movies, products, natural calamities, and government policies help in rapid decision-making. It helps in developing advanced natural text-to-speech systems with emotions. Health agencies can take rapid decision during epidemics and natural calamities. Emotions related to political events can help parties for improving their campaign. Detecting how people use emotion-bearing-words and metaphors to persuade and coerce others.

8. Conclusion and scope of future work

Public emotions present in Social media data offers unique challenges and opportunities for in decision-making in different domains. The major contribution of this research is to present that it is feasible to apply intelligent computational techniques for identification and classification of various types of emotions in texts. An effective technique for data collection and extraction of emotions in social media data has been presented through this paper. Important meaningful inferences are presented using multiple data sets. Classification of the dataset has been performed using machine learning techniques in two phases which provide better results in comparison to other approaches proposed by other authors. A comparison of corpus-based technique with proposed emotion model has been performed. Proposed emotion extraction framework classifies emotions present in Multilanguage data using different feature groups taken from publicly available lexical resources with improved accuracy. The combination of corpus-based features and emotion related features together improved performance in comparison to any type of feature group alone. This paper contributes in the field of Political election, Healthcare, and Sport analytic using social media which have a vast amount of information hidden in it.

The work presented in this paper can be pursued further in several domains. One of the tasks is to consider emotion intensity for classification. Explore the relation between emotion classes and emotion intensity. Content-based analysis of emotion data is yet another possible line of research. Data sets containing emoticons, stickers and other images with texts representing emotions can also be taken into consideration in future.

References

[1] G.L. Clore, A. Ortony, M.A. Foss, The psychological foundations of the affective lexicon, J. Pers. Soc. Psychol. 53 (1987) 751–766.
[2] C. Strapparava, A. Valitutti, WordNet-Affect: an affective extension of WordNet, in: Proceedings of 4th International Conference on Language Resources and Evaluation (LREC2004), Lisbon, Portugal, 2004, pp. 1083–1086.
[3] J. Bollen, H. Mao, X.-J. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1) (2011) 1–8.
[4] S. Aman, S. Szpakowicz, Identifying expressions of emotion in text, in: V. Matoǔsek, P. Mautner (Eds.), Text, Speech and Dialogue, Volume 4629 of Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, 2007, pp. 196–205.
[5] R. Mihalcea, C. Strapparava, Making Computers laugh: investigations in automatic humor recognition, in: Proceedings of the Joint Conference on Human Language Technology/Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, Canada, 2005, pp. 531–538.
[6] D. Ghazi, D. Inkpen, S. Szpakowicz, Hierarchical versus flat classification of emotions in text, in: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 140–146.
[7] V.K. Jain, S. Kumar, An effective approach to track levels of Influenza-A (H1N1) pandemic in India using Twitter, Procedia Comput. Sci. 70 (1) (2015) 801–807.
[8] P. Ekman, An argument for basic emotions, Cogn. Emot. 6 (1992) 169–200.
[9] S.S. Tomkins, Affect, Imagery, Consciousness. The Positive Affects, Springer, New York, 1962, pp. 23–48.
[10] C.E. Izard, Human Emotions, Plenum Press, New York, 1977, pp. 28–46.
[11] R. Plutchik, Emotion: A Psychoevolutionary Synthesis, Harper & Row, New York, 1980, pp. 19–48.
[12] A. Ortony, G.L. Clore, A. Collins, The Cognitive Structure of Emotions, Cambridge University Press, 1988, pp. 17–41.
[13] V. Raghavan, The Number of Rasa, Madras, 1940, pp. 20–45.
[14] J. Lei, Y. Rao, Q. Li, X. Quan, L. Wenyin, Towards building a social emotion detection system for online news, Future Gener. Comput. Syst. 37 (2014) 438–448.
[15] C.E. Osgood, G.J. Succi, P.H. Tannenbaum, The Measurement of Meaning, University of Illinois Press, Urbana, 1957, pp. 12–19.
[16] R. Jakobson, Linguistics and poetics, in: T. Sebeok (Ed.), Style in Language, MIT Press, Cambridge, MA, 1960, pp. 350–377.
[17] D. Watson, A. Tellegen, Towards a consensual structure of mood, Psychol. Bull. 98 (1985) 219–235.
[18] P.N. Johnson-Laird, K. Oatley, The language of emotions: an analysis of a semantic field, Cogn. Emot. 3 (2) (1989) 81–123.

[19] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, 1998, pp. 2–9.

[20] M.M. Bradley, P.J. Lang, Affective Norms for English Words (ANEW): Stimuli, Instruction Manual and Affective Ratings. Technical Report C-1, Gainesville, FL, The Center for Research in Psychophysiology, University of Florida, 1999, pp. 17–40.

[21] J. Kamps, M. Marx, Words with attitude, in: Proceedings of the First International Conference on Global WordNet, Mysore, India, 2002, pp. 6–9.

[22] H. Liu, H. Lieberman, T. Selker, A model of textual affect sensing using real-world knowledge, in: Proceedings of the International Conference on Intelligent User Interfaces, IUI 2003, Miami, FL, USA, 2003, pp. 125–132.

[23] J.R. Martin, P.R.R. White, the Language of Evaluation: Appraisal in English, Palgrave, London, 2005, pp. 26–38.

[24] C.O. Alm, D. Roth, R. Sproat, Emotions from text: machine learning for text-based emotion prediction, in: Proceedings of the Joint Conference on Human Language Technology/Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, Canada, 2005, pp. 579–586.

[25] G. Mishne, Experiments with mood classification in blog posts, in: Proceedings of the 1st Workshop on Stylistic Analysis of Text For Information Access (Style 2005), Brazil, 2005, pp. 321–327.

[26] C. Strapparava, A. Valitutti, O. Stock, The affective weight of lexicon, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, 2006, pp. 423–426.

[27] R.R. Mihalcea, H. Liu, A corpus-based approach to finding happiness, in: AAAI Spring Symposium on Computational Approaches to Weblogs, Technical Report SS-06-03, Stanford, CA, USA, 2006.

[28] L. Zhang, J. Barnden, R. Hendley, A. Wallington, Exploitation in affect detection in open-ended improvisational text, in: Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text, Sydney, Australia, 2006, pp. 47–54.

[29] J. Read, Using emoticons to reduce dependency in machine learning techniques for sentiment classification, in: Proceedings of the ACL 2005 Student Research Workshop, Ann Arbor, MI, USA, 2005, pp. 43–48.

[30] A. Neviarouskaya, H. Prendinger, M. Ishizuka, Analysis of affect expressed through the evolving language of online communication, in: Proceedings of the 12th International Conference on Intelligent User Interfaces, Honolulu, HI, USA, 2007, pp. 278–281.

[31] F.-R. Chaumartin, UPAR7: a knowledge-based system for headline sentiment tagging, in: Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics, 2007, pp. 422–522.

[32] D.T. Ho, T.H. Cao, A high-order hidden Markov model for emotion detection from textual data, in: Knowledge Management and Acquisition for Intelligent Systems, Springer, Berlin/Heidelberg, 2012, pp. 94–105.

[33] L. Dey, N. Afroz, R.P. Deb Nath, Emotion extraction from real time chat messenger, in: Proceedings of 3rd International Conference on Informatics, Electronics & Vision, Dhaka, Bangladesh, 2014, pp. 1–5.

[34] S. Shaheen, W. El-Hajj, H. Hajj, S. Elbassuoni, Emotion recognition from text based on automatically generated rules, in: 2014 IEEE International Conference on Data Mining Workshop, Shenzhen, China, 2014, pp. 383–392.

[35] J. Gordon, Comparative Geospatial Analysis of Twitter Sentiment Data During the 2008 and 2012 U.S. Presidential Elections. Master Thesis, 2013, pp. 23–41.

[36] S. Aman, S. Szpakowicz, Using Roget's Thesaurus for fine-grained emotion recognition, in: Proceedings of the Third International Joint Conference on Natural Language Processing, Hyderabad, India, 2008, pp. 296–302.

[37] D. Milne, C. Paris, H. Christensen, P. Batterham, B. ODea, We feel: taking the emotional pulse of the world, in: Proceedings of the 19th Triennial Congress of the International Ergonomics Association, Melbourne, Australia, 2015.

[38] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inf. Retrieval 2 (2008) 121–135.

[39] S.L. Fernandes, G.J. Bala, Fusion of sparse representation and dictionary matching for identifications of humans in uncontrolled environment, J. Comput. Biol. Med. 76 (2016) 215–237.

[40] A. Dogra, S. Agrawal, B. Goyal, C. Ahuja, N. Khandelwal, Color and grey scale fusion of osseous and vascular information, J. Comput. Sci. (DOI: http://dx.doi.org/10.1016/j.jocs.2016.09.003).

[41] S. Aman, Recognizing Emotions in Text, Master Thesis, School of Information Technology and Engineering, University of Ottawa, 2007.

[42] B. Ofoghi, M. Mann, K. Verspoor, Towards early discovery of salient health threats: a social media emotion classification technique, in: Proceedings of Pacific Symposium on Biocomputing (PSB), HI, US, 2016, pp. 504–515.

[43] M. Anjaria, R.M.R. Guddeti, Influence factor based opinion mining of Twitter data using supervised learning, in: 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), Bangalore, 2014, pp. 1–8.

[44] N. Li, D.D. Wu, Using text mining and sentiment analysis for online forums hotspot detection and forecast, Decis. Support Syst. 48 (2) (2010) 354–368.

[45] W. Wanga, D. Rothschild, S. Goel, A. Gelmana, Forecasting elections with non-representative polls, Int. J. Forecasting 31 (3) (2015) 980–991.

[46] B.O. Connor, R. Balasubramanyan, B.R. Routledge, N.A. Smith, From tweets to polls: linking text sentiment to public opinion time series, in: Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, 2010, pp. 122–129.

[47] A. Tumasjan, T. Sprenger, P. Sandner, I.I. Welpe, Predicting elections with Twitter: what 140 characters reveal about political sentiment, in: International AAAI Conference on Weblogs and Social Media, 2010, pp. 178–185.

[48] D.J. Hopkins, G. King, A method of automated nonparametric content analysis for social science, Am. J. Polit. Sci. 54 (1) (2010) 229–247.

[49] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, Election forecasts with twitter how 140 characters reflect the political landscape, Soc. Sci. Comput. Rev. 29 (4) (2011) 402–418.

[50] N.A. Diakopoulos, D.A. Shamma, Characterizing debate performance via aggregated twitter sentiment, in: Proceeding CHI'10 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2010, pp. 1195–1198.

[51] M. Choy, L.F.M. Cheong, M.N. Liak, K.P. Shung, A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction, 2011. Availible: [https://arxiv.org/ftp/arxiv/papers/1108/1108.5520.pdf].

[52] A. Bermingham, A.F. Smeaton, On using Twitter to monitor political sentiment and predict election results, in: Sentiment Analysis where AI meets Psychology (SAAIP) Workshop at the International Joint Conference for Natural Language Processing (IJCNLP), Chiang Mai, Thailand, 2011, pp. 145–152.

[53] P.T.A. Metaxas, How (not) to predict elections, in: IEEE Third International Conference on Social Computing (SocialCom), 2011, pp. 165–171.

[54] M.D. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, A. Flammini, F. Menczer, Political polarization on Twitter, in: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011, pp. 89–96.

[55] A.D. Maynard, A. Funk, Automatic detection of political opinions in Tweets, in: ESWC 2011 Workshops, Heraklion, Greece, May 29–30, 2011, pp. 88–99.

[56] A.O. Larsson, M. Hallvard, Studying political microblogging: Twitter users in the 2010 Swedish election campaign, New Media Soc. 14 (5) (2012) 729–747.

[57] M.C. Choy, US Presidential Election 2012 Prediction using Census Corrected Twitter Model, 2012, arXiv preprint arXiv:1211.0938.

[58] M. Skoric, N. Poor, P.E. Achananuparp, P. Lim, J. Jiang, Tweets and votes: a study of the 2011 Singapore general election, in: System Science (HICSS), 45th Hawaii International Conference, Hawaii, 2012, pp. 2583–2591.

[59] E. Tjong, K. Sang, J. Bos, Predicting the 2011 Dutch senate election results with Twitter, in: EACL 2012Workshop on Semantic Analysis in Social Networks, Avignon, France, April 23, 2012, pp. 53–60.

[60] J. Johnson, Twitter Bites and Romney: examining the rhetorical situation of the 2012 presidential election in 140 characters, J. Contemp. Rhetoric 2 (3) (2012) 54–64.

[61] F.B. Marquez, D. Gayo-Avello, M. Mendoza, B. Poblete, Opinion dynamics of elections in Twitter, in: Proceedings of the 2012 Eighth Latin American Web Congress, 2012, pp. 32–39.

[62] L. Shi, N. Agarwal, A. Agrawal, R. Garg, J. Spoelstra, Predicting US primary elections with Twitter, in: Workshop on Social Network and Social Media Analysis: Methods, Models and Applications (NIPS), 2012, pp. 1–8.

[63] J.M. Soler, F. Cuartero, M. Roblizo, Twitter as a tool for predicting elections results, in: Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on 26–29 Aug, 2012, pp. 1194–1200.

[64] D.T. Contractor, A. Faruquie, Understanding election candidate approval ratings using social media data, in: WWW 2013 Companion, Rio de Janeiro, Brazil, May 13–17, 2013, pp. 189–190.

[65] F.A. Nooralahzadeh, 2012 Presidential elections on Twitter – an analysis of how the US and French election were reflected in tweets, in: 19th International Conference on Control Systems and Computer Science (CSCS), 2013, pp. 240–246.

[66] H.D. Kim, M. Castellanos, M. Hsu, C.X. Zhai, T. Rietz, D. Diermeier, Mining causal topics in text data: iterative topic modeling with time series feedback, in: CIKM'13, San Francisco, CA, USA, 2013, pp. 885–890.

[67] A. Bakliwal, J. Foster, J.V. Puil, R. O'Brien, L. Tounsi, M. Hughes, Sentiment analysis of political tweets: towards an accurate classifier, in: Proceedings of the Workshop on Language in Social Media (LASM 2013), Atlanta, GA, 2013, pp. 49–58.

[68] N.A. Thapen, M.M. Ghanem, Towards passive political opinion polling using Twitter, in: BCS SGAI Workshop on Social Media Analysis, Cambridge, UK, 2013, pp. 19–34.

[69] M. Song, M.C. Kim, Y.K. Jeong, Analyzing the political landscape of 2012 Korean presidential election in Twitter, in: IEEE Intelligent Systems, 2014, pp. 1541–1672.

[70] C. Vaccari, A. Valeriani, P. Barberá, R. Bonneau, J.T. Jost, J. Nagler, J.A. Tucker, Political expression and action on social media: exploring the relationship between lower- and higher-threshold political activities among Twitter users in Italy, J. Comput. Mediat. Commun. 20 (2) (2015) 221–239.

[71] C.M. Chew, Pandemics in the Age of Twitter: A Content Analysis of the 2009 h1n1 Outbreak. Master's Thesis, University of Toronto, 2010, pp. 15–39.

[72] X. Hu, L. Tang, H. Liu, Enhancing accessibility of microblogging messages using semantic knowledge, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, New York, USA, 2011, pp. 2465–2468.

[73] V. Lampos, N. Cristianini, Tracking the flu pandemic by monitoring the social web, in: 2nd IAPR Workshop on Cognitive Information Processing (CIP 2010), IEEE Press, 2010, pp. 411–416.

[74] R. Chunara, J.R. Andrews, J.S. Brownstein, Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak, Am. J. Trop. Med. Hyg. 86 (1) (2012) 39–45.

[75] E. Aramaki, S. Maskawa, M. Morita, Twitter catches the flu: detecting influenza epidemics using Twitter, in: Proceedings of the Conference on Empirical

Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 1568–1576.

[76] A. Stewart, E. Diaz, Epidemic intelligence: for the crowd, by the crowd, in: Proceedings of the 12th International Conference on Web Engineering, ICWE'12, Springer-Verlag, Berlin/Heidelberg, 2012, pp. 504–505.

[77] T. Bodnar, V.C. Barclay, N. Ram, C.S. Tucker, M. Salathé, On the ground validation of online diagnosis with twitter and medical records, in: WWW Companion'14, 2014, pp. 651–656.

[78] J. Parker, Y. Wei, A. Yates, O. Frieder, N. Goharian, A framework for detecting public health trends with Twitter, in: Proceeding ASONAM'13 Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013, pp. 556–563.

[79] BurstMedia (2015) [http://www.postano.com/blog/how-social-media-is-changing-sports-marketing].

[80] X. Wang, Applying the integrative model of behavioral prediction and attitude functions in the context of social media use while viewing mediated sports, Comput. Human Behav. 29 (2013) 1538–1545.

[81] Y. Yu, X. Wang, World Cup 2014 in the Twitter World: a big data analysis of sentiments in U.S. sports fans' tweets, Comput. Human Behav. 48 (2015) 392–400.

[82] F. Godin, J. Zuallaert, B. Vandersmissen, W.D. Neve, R.V.D. Walle, Beating the bookmakers: leveraging statistics and Twitter microposts for predicting soccer results, in: KDD Workshop on Large-Scale Sports Analytics, Sydney, Australia, 2014.

[83] V. Radosavljevic, M. Grbovic, N. Djuric, N. Bhamidipati, Large-scale World Cup 2014 outcome prediction based on Tumblr posts, in: KDD Workshop on Large-Scale Sports Analytics, Sydney, Australia, 2014.

[84] S. Sinha, C. Dyer, K. Gimpel, N. Smith, Predicting the NFL using Twitter, 2013, arXiv Preprint Retrieved from http://arxiv.org/abs/1310.6998 arXiv:1310.6998, 1-11.

[85] D. Lock, D.D. Nettleton, Using random forests to estimate win probability before each play of an NFL game, J. Quant. Anal. Sports 10 (2) (2014) 9–15.

[86] N. UzZaman, R. Blanco, M. Matthews, TwitterPaul: Extracting and Aggregating Twitter Predictions. Artificial Intelligence; Physics and Society. 2014. Available: [http://arxiv.org/abs/1211.6496].

[87] D.M. Blei, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[88] Y. Bao, C. Quan, L. Wang, F. Ren, The role of pre-processing in Twitter sentiment analysis, in: ICIC, Taiyuan, China, 2014, pp. 615–624.

[89] Hindi WordNet, IIT Mumbai, Available: [http://www.cfilt.iitb.ac.in/∼wordnet/wn.old/].

[90] Senti Wordnet, Available: [http://sentiwordnet.isti.cnr.it/].

[91] M. Lui, B. Timothy, Accurate language identification of Twitter messages, in: Proceedings of the EACL 2014 Workshop on Language Analysis in Social Media, Gothenburg, Sweden, 2014, pp. 17–25.

[92] M. Lui, B. Timothy, langid.py: An Offtheshelf Language Identification Tool, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Demo Session, Jeju, Republic of Korea, 2012, pp. 25–30.

Vinay Kumar Jain received his Bachelor's Degree in 2009 from Rajiv Gandhi Proudyogiki Vishwavidyala, Bhopal, India and received his Master's Degree from Jaypee University of Engineering and Technology, India in 2012. Now, he is pursuing his Ph.D. degree from Jaypee University of Engineering and Technology, Guna, M.P., India.

Shishir Kumar in working as Professor the Department of Computer Science and Engineering at Jaypee University of Engineering and Technology, Guna, M.P., India. He has earned Ph.D. in Computer Science in 2005. He has 14 years of teaching and research experience.

Steven Fernandes is member of Core Research Group, Karnataka Government Research Centre of Sahyadri College of Engineering and Management, Mangalore, Karnataka. He has received Young Scientist Award by Vision Group on Science and Technology, Government of Karnataka, India in the year 2014 and also received grant from The Institution of Engineers (India), Kolkata, India. He completed his B.E. (Electronics and Communication Engineering) with Distinction from Visvesvaraya Technological University, Belagavi, Karnataka and M.Tech. (Microelectronics) with Distinction from Manipal University, Manipal, Karnataka. His Ph.D. work "Match Composite Sketch with Drone Images" has received patent notification (Patent Application Number: 2983/CHE/2015) from Government of India, Controller General of Patents, Designs & Trade Marks. He has 5 years of industry experience working at STMicroelectronics Pvt. Ltd. and Perform Group Pvt. Ltd. He has published several papers in peer-reviewed International Journals having Thomson Reuters Web of Science Impact Factor and IEEE, Springer, Elsevier International Conferences. He is also serving has reviewer and guest editor for several Science Citation Indexed and Scopus Indexed International Journals.