TOM: Twitter opinion mining framework using hybrid classification scheme

Article *in* Decision Support Systems · January 2014

DOI: 10.1016/j.dss.2013.09.004

CITATIONS
147

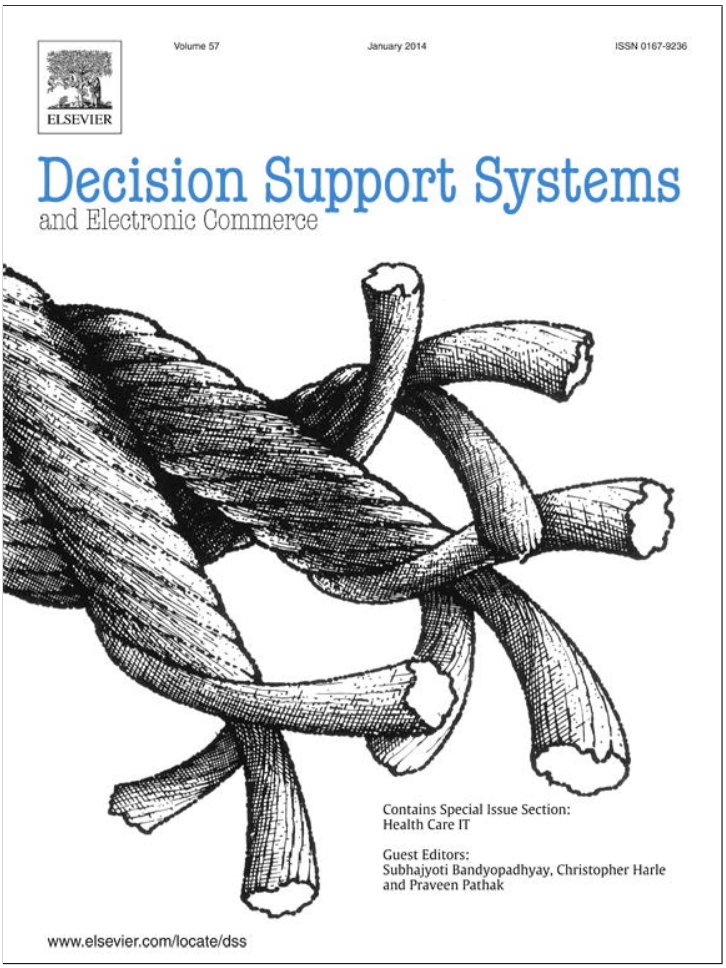
READS
1,748

3 authors:

Farhan Hassan Khan
National University of Sciences and Technology
44 PUBLICATIONS **420** CITATIONS
SEE PROFILE

Saba Bashir
National University of Sciences and Technology
38 PUBLICATIONS **432** CITATIONS
SEE PROFILE

Usman Qamar
National University of Sciences and Technology
162 PUBLICATIONS **643** CITATIONS
SEE PROFILE



Decision Support Systems 57 (2014) 245–257

Contents lists available at ScienceDirect

Decision Support Systems

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

TOM: Twitter opinion mining framework using hybrid classification scheme

Farhan Hassan Khan, Saba Bashir *, Usman Qamar

Computer Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 20 February 2013
Received in revised form 30 July 2013
Accepted 11 September 2013
Available online 21 September 2013

Keywords:
Twitter
Sentiment analysis
Classification
SentiWordNet
Social network analysis
Data sparsity

ABSTRACT

Twitter has become one of the most popular micro-blogging platform recently. Millions of users can share their thoughts and opinions about different aspects and events on the micro-blogging platform. Therefore, Twitter is considered as a rich source of information for decision making and sentiment analysis. Sentiment analysis refers to a classification problem where the main focus is to predict the polarity of words and then classify them into positive and negative feelings with the aim of identifying attitude and opinions that are expressed in any form or language. Sentiment analysis over Twitter offers organisations a fast and effective way to monitor the publics' feelings towards their brand, business, directors, etc. A wide range of features and methods for training sentiment classifiers for Twitter datasets have been researched in recent years with varying results. The primary issues in previous techniques are classification accuracy, data sparsity and sarcasm, as they incorrectly classify most of the tweets with a very high percentage of tweets incorrectly classified as neutral. This research paper focuses on these problems and presents an algorithm for twitter feeds classification based on a hybrid approach. The proposed method includes various pre-processing steps before feeding the text to the classifier. Experimental results show that the proposed technique overcomes the previous limitations and achieves higher accuracy when compared to similar techniques.

1. Introduction

The emergence of social media has given web users a venue for expressing and sharing their thoughts and opinions on different topics and events. Twitter, with nearly 600 million users and over 250 million messages per day, has quickly become a gold mine for organisations to monitor their reputation and brands by extracting and analysing the sentiment of the Tweets posted by the public about them, their markets, and competitors.

Sentiment analysis over Twitter data and other similar micro-blogs faces several new challenges due to the typical short length and irregular structure of such content. The following are some challenges faced in sentiment analysis of Twitter feeds.

- Named Entity Recognition (NER) — NER is the method of extracting entities such as people, organisation and locations from twitter corpus.
- Anaphora Resolution — The process of resolving the problem of what a pronoun or noun phrase refers to. "We had a lavish dinner and went for a walk, it was awful". What does "It" refer to?
- Parsing — The process of identifying the subject and object of the sentence. The verb and adjective are referring to what?
- Sarcasm — What does a verb actually stand for? Does 'bad' mean bad or good?
- Sparsity — Insufficient data or very few useful labels in the training set.
- Twitter abbreviations, poor spellings, poor punctuation, poor grammar, incomplete sentences.
- The accuracy of tweets classification as compared to human judgments.

Fig. 1 presents a generic framework of sentiment analysis where a sentiment engine receives feedback (data) from different channels and then a unique algorithm categorizes (positive/negative) them by assigning scores. The results can be used to draw various types of graphs which are presented in the dashboard. These results can be used to find out the overall feelings towards a particular person, product or service.

This research paper presents a technique for text mining of Twitter feeds in real time and sentiment analysis using three-way classification by investigating the sentiment intensity. The main focus is on improving the accuracy and solving the data sparsity issue in tweet classification, effectively reducing the number of tweets classified as neutral.

1.1. Motivation

Micro-blogging website Twitter has evolved to become a source of rich and varied information. This is due to nature of micro-blogs on which people post real time messages about their opinions on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in daily life. In fact, companies manufacturing such products have started to poll these tweets to get a sense of general sentiment for their product. Many times these companies study user reactions and reply to users on twitter. However, the sheer

* Corresponding author.
E-mail addresses: mrfarhankhan@gmail.com (F.H. Khan), saba.bashir3000@gmail.com (S. Bashir), usmanq@ceme.nust.edu.pk (U. Qamar).

0167-9236/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2013.09.004

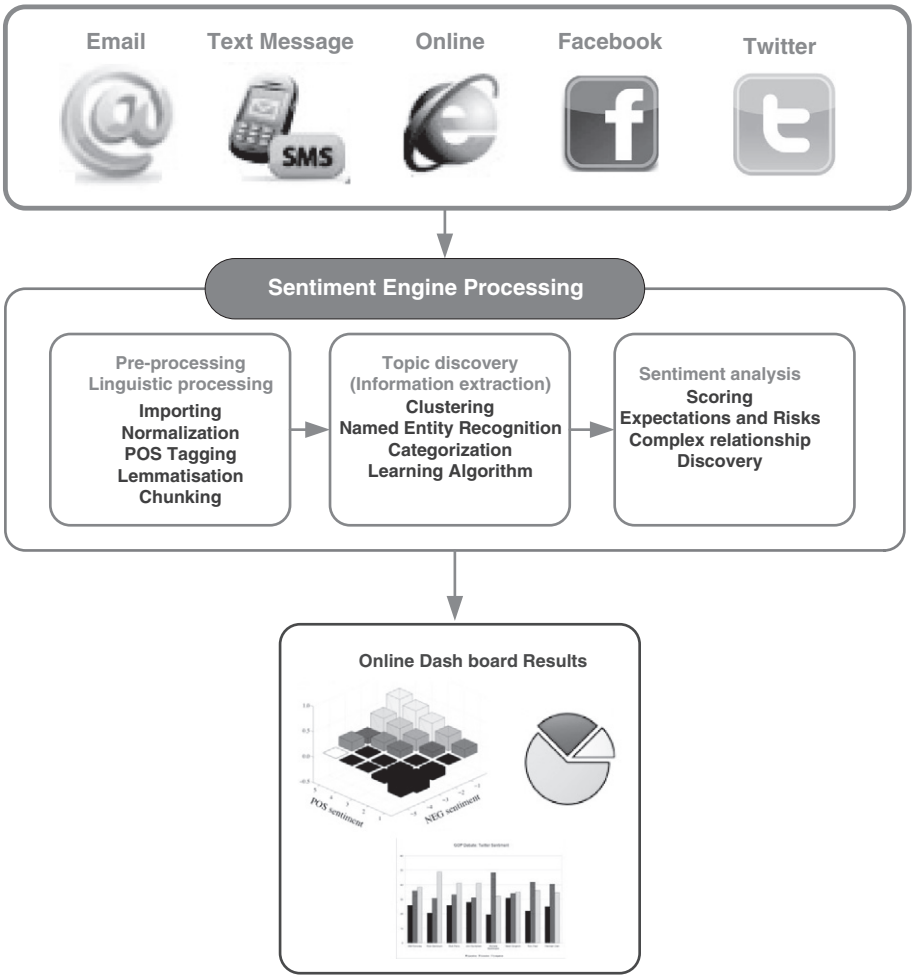


Fig. 1. Generic sentiment analysis framework.

vastness of data makes it very difficult to analyse and grasp this data. Therefore, it is necessary to automate the process of analysing the twitter's data. In order to complete this task, there is an immense need to automatically classify the twitter's tweets as positive, negative or neutral in real time.

The main contributions of this paper can be summarised as follows:

- Introduces and implements a hybrid approach for determining the sentiment of each tweet
- Demonstrates the value of pre-possessing data using detection and analysis of slangs/abbreviations, lemmatization, correction and stop words removal
- Tests the accuracy of sentiment identification on 6 Twitter datasets, and produces an average harmonic mean of 83.3% and accuracy of 85.7%, with 85.3% precision and 82.2% recall
- Resolves the data sparsity issue using domain independent techniques.
- Comparison with other techniques to prove the effectiveness of the proposed hybrid approach.

The structure of the paper is described as follows. Section 2 outlines the recent related work. Section 3 introduces the proposed technique while Section 4 gives an overview of the results. Finally Section 5 summarises the work which has been done.

2. Literature review

There are multiple text mining techniques used to mine the twitter feeds.

Cui, A. et al. [1] showed that sentiment analysis of tweets is a challenging task due to multilingual and informal messages. The paper tackles this problem by analysis of emotion tokens. Emotion is the mood of a person depicted from the words in the tweet. Emotion can be sad, happy, angry, etc. The proposed approach has two steps. First, emotion tokens are extracted from the message. Second, graph propagation algorithm plots the tokens at different polarities. Finally, sentiment analysis algorithm analyses and classifies these emotion tokens. The results show that emotion tokens are a great approach towards semantic analysis of any natural language where lexicons are built independent of different time domains. The technical issues in the proposed approach are: less accuracy of sentiment analysis, difficulty in tackling sentiment analysis of Twitter stream for longer time and weak emotion representation.

Bifet, A. and Frank, E. [2] proposed a data mining technique used for sentiment knowledge in twitter data streams. The proposed algorithm focuses on classification of data streams and performs sentiment analysis in real time. The evaluation of results is verified using a sliding window kappa statistics that works for constantly changing data streams. Only a small number of tweets (177 negative and 182 positive) were used to test the accuracy. This is a very small number of tweets to make any judgment about the proposed technique. Only tweets containing an emoticon were considered; which is also a very small portion of overall tweets. The paper uses a balanced dataset which is not a sample of real-time Twitter stream which is normally unbalanced.

Bifet, A., Holmes, G., and Pfahringer, B. [3] discussed the handling of tweets in real-time. The research paper introduced a system, MOA-TweetReader, which processes the tweets in real-time despite their

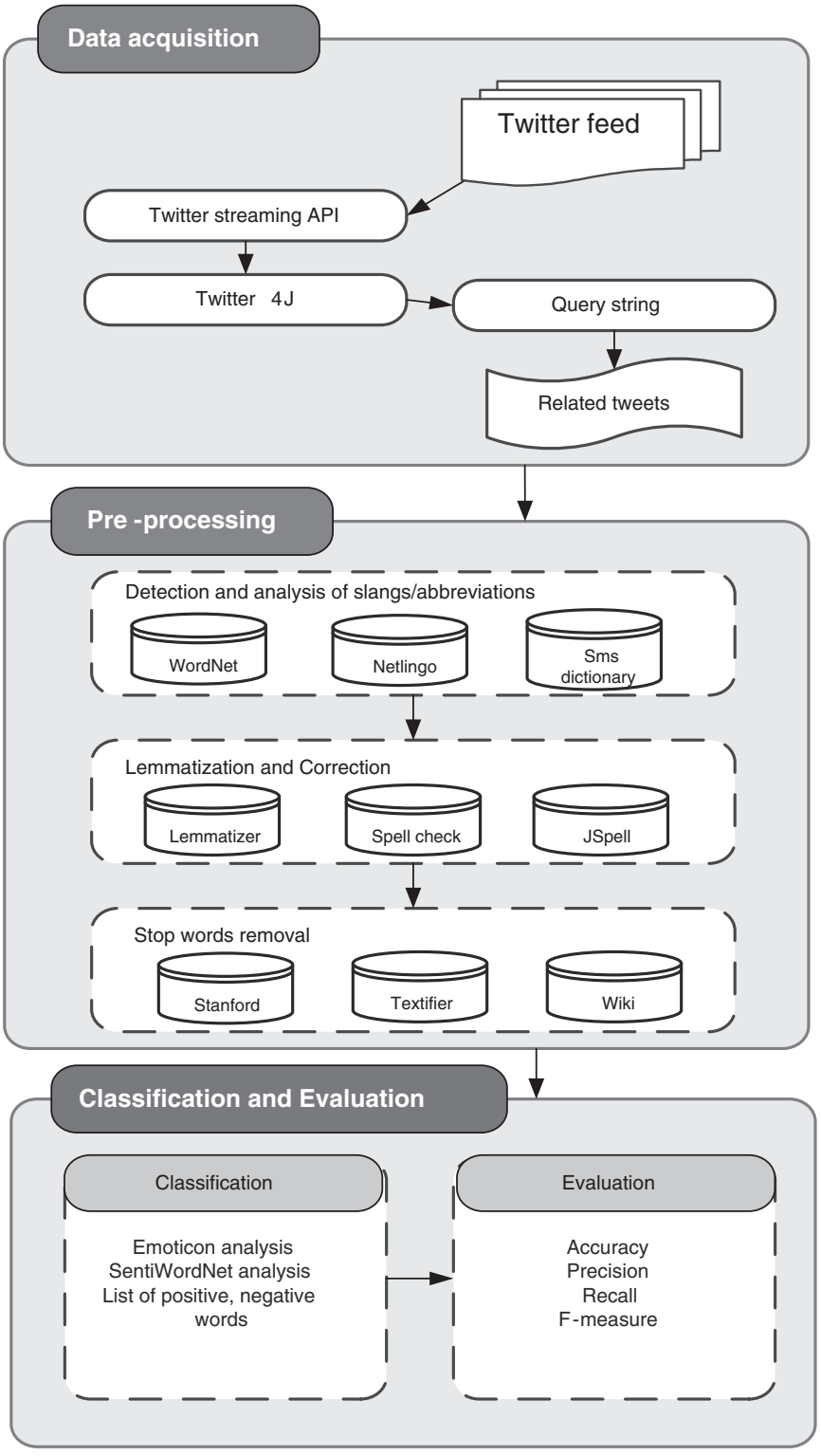


Fig. 2. Detailed architecture of the proposed TOM framework.

dynamic nature. The system performs two functions: First, it detects the changes in term frequencies and second, it performs sentiment analysis in real-time. Some applications of the proposed framework have also been discussed in frequent item mining and sentiment analysis. The paper shows a correlation between the twitter sentiments and the Toyota crisis and successfully claims that the MOA-TweetReader tool could have identified the crisis coming. The authors only use positive/negative classes for sentiment classification. There is a possibility that a tweet may be neutral, which is not considered in this paper. A small set of emoticons (5 positive and 3 negative) are used, which may not be enough to discover all the emoticons in the tweets.

Ye, S. and Wu, S. F. [4] discovered a message propagation pattern using Twitter. The evaluation is based on examining different social influences and their effects such as stabilities, correlations and assessments. An important feature of this research is the identification of popular tweets. However, the authors do not explain the criteria for a

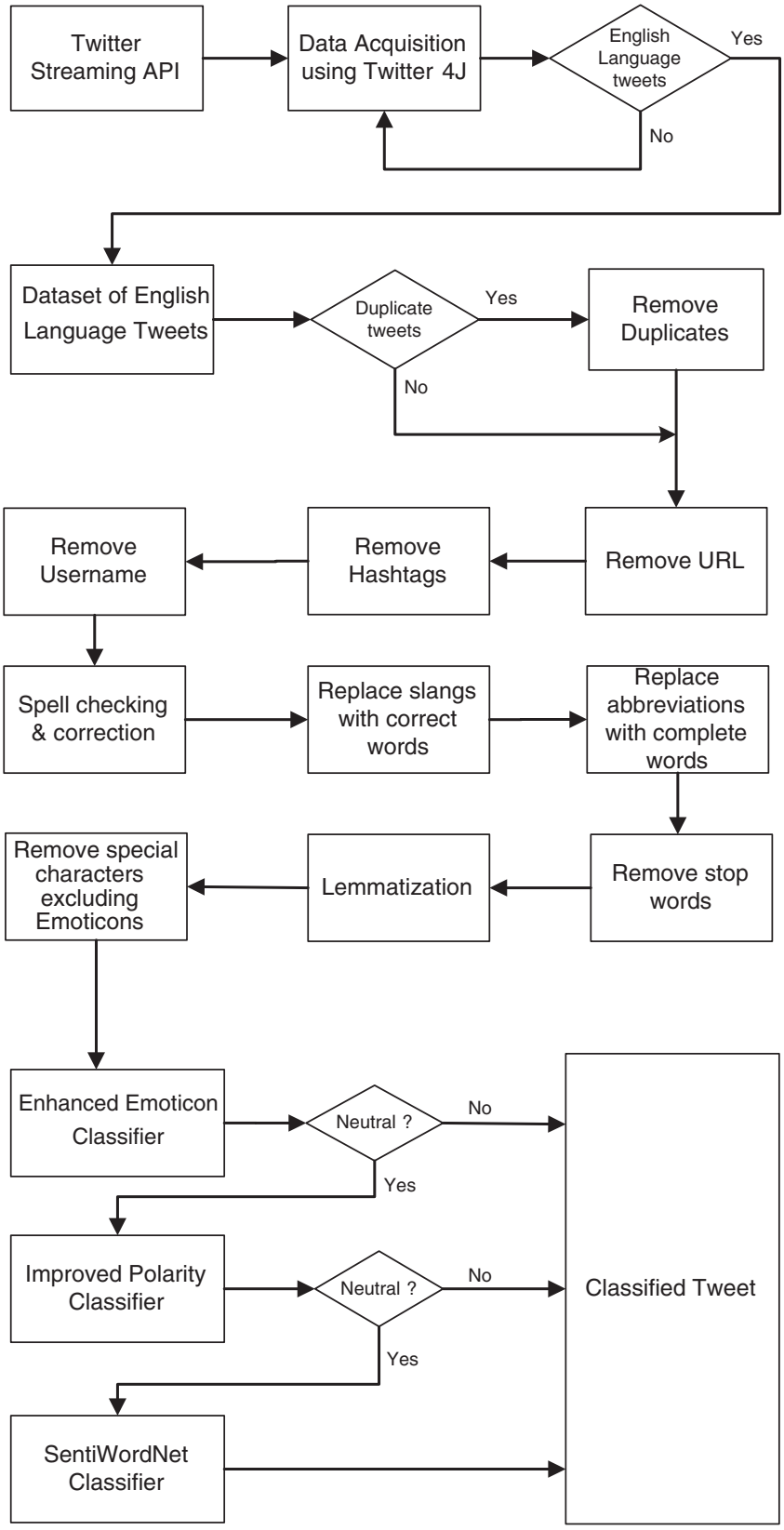


Fig. 3. Flow chart of proposed TOM framework.

tweet being popular. The paper also does not explain how it identifies the virus/worms posting tweets for marketing purposes. Social influence is categorized into reply influence and re-tweet influence. There is no discussion about polarity of the influence. The analysis of results is a great achievement towards systematic measurement and investigations on OSNs.

Argamon, S. et al. [5] used a supervised learning algorithm for determining complex sentiment-related attributes. These attributes

```
Begin
    Input: QueryString
    Until the data is retrieved from Twitter Streaming API, Do
        Filter English Language Tweets
        Remove Duplicates
        FOR each tweet, Do
                Procedure Pre-process (tweet)
                        Remove URL
                        Remove Hashtags
                        Remove Username
                        Spell Check & Correction
                        Replace Slangs
                        Replace Abbreviations
                        Remove Stop Words
                        Lemmatization
                        Remove Special Characters
                End Procedure
                Procedure Classification (Refined tweet)
                        Classify refined tweet using Enhanced Emoticon Classifier
                        IF tweet is classified NEUTRAL
                                Classify refined tweet using Enhanced Polarity Classifier
                        END IF
                        IF tweet is classified NEUTRAL
                                Classify refined tweet using SentiWordNet Classifier
                        END IF
                        Write the classification result to file
                END Procedure
    End Until
End
```

Fig. 4. Proposed PCA for sentiment analysis.

are classified as attitude type and force. The WordNet glosses are used for the implementation of supervised learning algorithm which classifies them into four force levels and eleven of attitude levels. The effectiveness of the algorithm is shown through experimental results. The results when averaged show that the Naïve Bayes algorithm is the best among the lot. SVM also dominates the results. The proposed algorithm is well suited where lexicons need to be generated from the scratch.

Fu, X. et al. [6] presented a method for semantic extraction using information theoretic co-clustering. The proposed algorithm is based on implicit associations within evaluated features, within evaluated semantic words and between evaluated features and semantic words. A feature semantic word matrix represents the co-occurrence relationships of feature words and semantic words. Then, co-clustering algorithm is applied on this matrix for clustering of these evaluated features. The dataset used for testing is in Chinese language. There are no details about how the manual analysis of the online review was conducted to ensure that the analysis was not biased. The effectiveness of the proposed technique is demonstrated in experimental results which show 78% accuracy.

Nagy, A. and Stamberger, J. [7] proposed a technique to efficiently identify the sentiment from disaster micro-blogs. The paper presents a technique for crowd sentiment detection from informative messages that are crafted for the crowd. This technique is useful to detect the sentiment during disaster and crises. Comparing against the Bayesian network, the proposed technique achieved 27% better performance. Two main approaches used are list based and classification based. The ensemble based technique is used for tweet categorization. SentiWordNet 3.0, list of emoticons, sentiment based dictionary and list of out of vocabulary words is used for sentiment detection from tweet. The emoticon is a pictorial representation of a facial expression depicting the mood of a person as angry, sad, normal or happy. The basic advantage of the proposed technique is that it is generic enough to detect the emotional pulse of people during disasters and crises. It can also detect and analyse the sentiment of a given tweet automatically. The technique can be applied to disaster data and with minimal tweaking and customization, the sentiments of new events and disasters can be identified. Limitations of the technique include limited ability to expand the initial seeds and continuous maintenance of lists.

Montejo-Raez, A. et al. [8] proposed an unsupervised approach for sentiment polarity detection from twitter tweets. The polarity scores are calculated from SentiWordNet and random walk algorithm is used to calculate the weights from the tweet. The proposed algorithm has comparable performance with SVM algorithm. The benefit of the proposed technique is that there is no need of training corpus as required in supervised learning techniques and there is no dependency on the model domain. The limitations include handling of negation, manual labelling process for certain tweets and facing flaws in calculation of final polarity score.

Ortega, R. et al. [9] proposed a technique with three phases; pre-processing, polarity identification and classification. WordNet and SentiWordNet based approach is used for the purpose of polarity

Table 1
Positive emoticons sample set.

Emoticon	Meaning
:-) :) :o) :] :3 :c) :> =] 8) =) :} :^) :)	Smiley or happy face
:-D :D 8-D 8D x-D xD X-D XD =-D =D =-3 =3 B^D	Laughing, big grin
:-))	Very happy or double chin

Table 2
Negative emoticons sample set.

Emoticon	Meaning
:-) :) :o) :] :3 :c) :> =] 8) =) :} :^) :)	Smiley or happy face
:-D :D 8-D 8D x-D xD X-D XD =-D =D =-3 =3 B^D	Laughing, big grin
:-))	Very happy or double chin

Table 3
Positive words sample.

Accurate	Beautiful	Capable	Decent	Ease	Faith
Gain	Happy	Immaculate	Joy	Keen	Lavish
Majestic	Neat	Optimal	Patriot	Quick	Rapid
Savvy	Thank	Unity	Valuable	Welcome	Youthful

detection and rule-based classification is performed. Good classification results are achieved for twitter data. However, there are no details about the dataset being used and the implementation of referenced algorithms. There is no comparison of results with the existing techniques to prove the effectiveness of the proposed research.

Bravo-Marquez, F., Mendoza, M. and Poblete, B. [10] combine the existing techniques for opinion strength, emotion and polarity prediction. The results show a great improvement in the classification process. The proposed method is not clearly presented. The authors do not discuss the differences and advantages of the proposed method when comparing with other existing methods.

Kim, J., et al. [11] proposed a collaborative filtering based model for predicting sentiment in Twitter. Two Twitter datasets were used for evaluation of the proposed model. The results show the effectiveness of the proposed approach for sentiment prediction and provide a solution for data sparsity problem. Few existing approaches have been used but there is no clear presentation of proposed framework which makes it very difficult to follow the method.

Machedon, R., Rand, W. and Joshi, Y. [12] defined that social media messages are classified into three categories; informative, persuasive and transformative. The tweet data is collected for 65 music bands where each tweet was labelled by human. The authors do not explain if there was any checking process performed to reach a good set of judgments. The results show that the proposed method is effective for the 'informative' category only.

Balahur, A. [13] presents a sentiment analysis technique for Twitter data. Training models are generated using the proposed method and the results show good classification performance. Minimal linguistic processing is applied in order to support multilingual datasets. Tweets contain a lot of slangs and abbreviations which would not be interpreted in this case. There is no comparison of results with other techniques in order to judge the effectiveness of the proposed technique.

All the techniques discussed in this section have some advantages and limitations. Hence a comprehensive technique is still needed to overcome their limitations.

3. The proposed framework (TOM)

The TOM framework applies a variant of techniques for Twitter feed analysis and classification. This involves pre-processing steps and a hybrid scheme of classification algorithms. Pre-processing steps include: removal of URLs, hash-tags, username & special characters; spelling correction using a dictionary; substitution of abbreviations & slangs with expansions, lemmatization and stop words removal. The proposed classification algorithm incorporates a hybrid scheme using an enhanced form of emoticon analysis [14], SentiWordNet analysis [15] and an improved polarity classifier using list of positive/negative words [16,21,22].

Table 4
Negative Words Sample.

Abnormal	Bad	Cancer	Danger	Enemy	Fake
Garbage	Hatred	Idiot	Jealous	Kill	Lazy
Malicious	Nervous	Obscure	Pain	Quarrel	Rage
Sad	Taunt	Upset	Vague	Weird	Zombie

Table 5
Example Tweets.

Sentiment	Query	Tweet
Positive	Imran Khan	RT@PTIMalaysia: people of Pakistan love imran khan
Negative	Tom Cruise	#Celebrity #Headline: Tom Cruise's "Jack Reacher" fails to reach audiences
Neutral	America	I want to travel around America. #travel #america

Previous research [17–19] on Twitter sentiment analysis present different techniques for text classification where the classifier performs a classification task on the basis of trained data set and machine learning (supervised) algorithms. Classifiers are trained on labelled corpora where the tweets are classified as positive and negative using emoticon analysis and other features like unigrams, part of speech tags and bigrams etc. Although these are good methods for tweet classification, they pose several challenges. The main challenges are: classification accuracy, sarcasm and data sparsity problem, as they incorrectly classify most of the tweets. The reason behind these problems is use of slangs and other shorthand grammars due to the limit of tweet message (140 characters). Major issues in supervised learning techniques are the availability of trained dataset and determining the structure of learned function. In contrast, unsupervised learning algorithms are a good option as they do not require any training data. Moreover, their goal depends only on situations and they are not mathematically well-defined.

The main goal of this research is to improve the accuracy of text classification and resolve the data sparsity issues. The core idea is to preprocess the raw data and perform different transformations to remove the slangs, grammatical mistakes, abbreviations and other noise and then feed it to the classifier. The TOM system is able to test the data streams from twitter streaming API in real-time continuously. The tweets obtained from these data streams are used as input items. The proposed system is basically composed of three main modules. The first module is data acquisition, a process of obtaining twitter feeds from OSN; the second module performs pre-processing and transforms the tweets containing real valued features or arbitrary components and refines them into a stream pattern that can be easily used for subsequent analysis. The last component applies different classification techniques in a pipelined way which classifies the tweets into positive, negative or neutral. The proposed framework is shown in Fig. 2 and the flowchart of the proposed framework is given in Fig. 3.

3.1. Proposed data acquisition technique

The fundamental purpose of data acquisition module is to obtain the Twitter feeds with sparse features in continuous fashion. The Twitter streaming API allows real time access to publicly available data on OSN. Twitter4J [20] library has been used for this purpose. The library was configured to extract only English language tweets. The tweets serve as input to pre-processing module and then they are further classified as positive, negative or neutral.

Table 6
Sample datasets.

	Query string	No of tweets analysed
Dataset 1	Imran Khan	99
Dataset 2	Nawaz Sharif	105
Dataset 3	Dhoni	100
Dataset 4	Tom Cruise	300
Dataset 5	Pakistan	512
Dataset 6	America	1000

Table 7
Confusion matrix.

		Predicted class		
		A	B	C
Known class	A	tpA	eAB	eAC
	B	eBA	tpB	eBC
	C	eCA	eCB	tpC

3.2. Proposed pre-processing steps

The pre-processing module involves performing intensive processing steps at each tweet individually and then passes each refined tweet to the classifier. This consists of following steps:

- Look up for meaning of each word in three English dictionaries (WordNet/SpellCheck/JSpell). The words that are not found illustrate that they are either slangs or abbreviations. For example, the tweet "@xyz u and Jane are gud friends". "u" and "gud" will not return any meaning.
- Abbreviations and/or shorthand notations will be replaced by expansions. Netlingo and sms dictionary are used for this purpose. Our example tweet will now be represented as, "@xyz you and Jane are good friends".
- The next step is to apply lemmatization. Lemmatization is used to stem the words and apply corrections. For example, when 'happiness' is stemmed to 'happi'.
- Apply spell checking of the tweet in order to correct the effects of the lemmatizer. This step feeds the remaining words in the spell checker and substitute with the best match. We have used Jazzy Spell Checker, JSpell and Snow ball for spell checking. For instance, 'happi' is corrected to 'happy'.
- Identify and remove the stop words. Stanford, Wiki and Textifier are used to identify the stop words which are then simply stripped from the tweet under process.
- Identify presence of URL using a regular expression and remove all the URLs from the tweet.
- Remove all the private usernames identified by @user and the hashtags identified by the # symbol.
- Lastly, remove all the special characters excluding the emoticons.
- The refined tweets are then classified using hybrid classification scheme.

3.3. Proposed Polarity Classification Algorithm and evaluation procedure

The proposed Polarity Classification Algorithm (PCA) in TOM framework classifies twitter feeds on the basis of

- Enhanced Emoticon Classifier (EEC)
- Improved Polarity Classifier (IPC)
- SentiWordNet Classifier (SWNC)

In EEC, classification is done on the basis of emoticons. It uses regular expressions to detect presence of emoticons which are then classified into positive or negative using a rich set of emoticons which are manually tagged as positive or negative. IPC uses a list of positive and negative words which are actually two text files that include positive and negative words respectively. SWNC is based on classification of tweets using SentiWordNet dictionary to check the sentiments for each word in the tweet.

Fig. 4 gives the detailed algorithm used for twitter feeds classification. Firstly, each tweet is pre-processed using pre-process procedure and then classification is performed at each refined tweet as defined in classification procedure. Finally, the output is generated in the form of positive, negative or neutral labelled tweets.

Let T be a set of tweets t defined as:

$$T = \{t_1,\ t_2,\ \ldots\ldots,\ t_n\}.$$

Table 8
Dataset 1 experiment and results.

Dataset 1		Confusion matrices			Results			
		Positive	Negative	Neutral	Precision	Recall	F–measure	Accuracy
Proposed	Positive	39	5	2	95.12%	84.78%	89.66%	88.89%
	Negative	2	38	0	84.44%	95.00%	89.41%	
	Neutral	0	2	11	84.62%	84.62%	84.62%	
EEC	Positive	8	0	38	100.00%	17.39%	29.63%	21.21%
	Negative	0	0	40	0.00%	0.00%	0.00%	
	Neutral	0	0	13	14.29%	100.00%	25.00%	
IPC	Positive	19	9	18	100.00%	41.30%	58.46%	65.66%
	Negative	0	33	7	78.57%	82.50%	80.49%	
	Neutral	0	0	13	34.21%	100.00%	50.98%	
SWNC	Positive	27	17	2	62.79%	58.70%	60.67%	61.62%
	Negative	16	23	1	54.76%	57.50%	56.10%	
	Neutral	0	2	11	78.57%	84.62%	81.48%	

Table 9
Dataset 2 experiment and results.

Dataset 2		Confusion matrices			Results			
		Positive	Negative	Neutral	Precision	Recall	F–measure	Accuracy
Proposed	Positive	38	3	0	82.61%	92.68%	87.36%	82.86%
	Negative	8	32	2	80.00%	76.19%	78.05%	
	Neutral	0	5	17	89.47%	77.27%	82.93%	
EEC	Positive	4	0	37	100.00%	9.76%	17.78%	23.81%
	Negative	0	0	42	0.00%	0.00%	0.00%	
	Neutral	0	1	21	21.00%	95.45%	34.43%	
IPC	Positive	16	3	22	76.19%	39.02%	51.61%	55.24%
	Negative	5	22	15	81.48%	52.38%	63.77%	
	Neutral	0	2	20	35.09%	90.91%	50.63%	
SWNC	Positive	30	10	1	65.22%	73.17%	68.97%	68.57%
	Negative	14	25	3	65.79%	59.52%	62.50%	
	Neutral	2	3	17	80.95%	77.27%	79.07%	

Let W be a set of words w in each tweet t defined as:

$$W = \{w_1,\ w_2,\ \ldots\ldots,\ w_m\}.$$

Score S is calculated as follows:

$$\text{Score} = \sum_{i=1}^{n}\sum_{j=1}^{m} S_{t_i w_j}.$$

where S is the sentiment score calculated for each word in the tweet.

We will use following three score calculations for the final classification of the tweet.

3.3.1. EEC score calculation

Emoticons are domain and language independent. They are used very sparingly and they constitute a very small portion of the text. As observed by Read [14], only 2.435% of downloaded Usenet articles contained a wink emoticon. EEC is very effective when the emoticons are present in the data. Read [14] achieved 70% accuracy for article extracts from emoticon dataset. EEC is not effective when classifying

Table 10
Dataset 3 experiment and results.

Dataset 3		Confusion matrices			Results			
		Positive	Negative	Neutral	Precision	Recall	F–measure	Accuracy
Proposed	Positive	35	3	0	79.55%	92.11%	85.37%	86.00%
	Negative	7	46	1	92.00%	85.19%	88.46%	
	Neutral	2	1	5	83.33%	62.50%	71.43%	
EEC	Positive	13	0	25	100.00%	34.21%	50.98%	23.00%
	Negative	0	2	52	100.00%	3.70%	7.14%	
	Neutral	0	0	8	9.41%	100.00%	17.20%	
IPC	Positive	14	3	21	87.50%	36.84%	51.85%	48.00%
	Negative	2	27	25	87.10%	50.00%	63.53%	
	Neutral	0	1	7	13.21%	87.50%	22.95%	
SWNC	Positive	24	10	4	58.54%	63.16%	60.76%	67.00%
	Negative	15	38	1	77.55%	70.37%	73.79%	
	Neutral	2	1	5	50.00%	62.50%	55.56%	

Table 11
Dataset 4 experiment and results.

Dataset 4		Confusion matrices			Results			
		Positive	Negative	Neutral	Precision	Recall	F-measure	Accuracy
Proposed	Positive	188	10	0	87.44%	94.95%	91.04%	85.00%
	Negative	27	67	0	78.82%	71.28%	74.86%	
	Neutral	0	8	0	0.00%	0.00%	0.00%	
EEC	Positive	16	0	182	100.00%	8.08%	14.95%	8.00%
	Negative	0	1	93	50.00%	1.06%	2.08%	
	Neutral	0	1	7	2.48%	87.50%	4.83%	
IPC	Positive	78	7	113	96.30%	39.39%	55.91%	44.33%
	Negative	3	48	43	85.71%	51.06%	64.00%	
	Neutral	0	1	7	4.29%	87.50%	8.19%	
SWNC	Positive	180	18	0	77.92%	90.91%	83.92%	75.00%
	Negative	49	45	0	65.22%	47.87%	55.21%	
	Neutral	2	6	0	0.00%	0.00%	0.00%	

datasets that do not contain any emoticon. From these points we come to a conclusion that EEC should be used when an emoticon is present in the tweet. Otherwise, some other classifier may be used for classification.

An emoticon is identified using a regular expression. The emoticon classification is based on sets of positive & negative emoticons. EEC is an enhancement of the technique proposed by Read [14]. Read used 11 trained emoticons whereas we have used a total of 145 emoticons; 70 of which are tagged positive and 75 are tagged as negative. The sample sets of positive and negative emoticons are given in Tables 1 and 2 respectively. Furthermore, we feed pre-processed text for classification. If the emoticon is found in the positive set then it is declared as positive. The emoticon is declared negative if it is found in negative set. If the emoticon is not found in both the sets, we declare it as neutral. Total positive and negative emoticons are counted and the sum is calculated. The refined tweet is assigned a score of 1 if the sum is greater than zero. A score of −1 is assigned if the sum is less than zero. A score of zero indicates that the calculated sum is zero.

Table 12
Dataset 5 experiment and results.

Dataset 5		Confusion matrices			Results			
		Positive	Negative	Neutral	Precision	Recall	F-measure	Accuracy
Proposed	Positive	177	16	2	80.45%	90.77%	85.30%	85.55%
	Negative	37	220	4	89.80%	84.29%	86.96%	
	Neutral	6	9	41	87.23%	73.21%	79.61%	
EEC	Positive	21	1	173	72.41%	10.77%	18.75%	16.02%
	Negative	8	6	247	75.00%	2.30%	4.46%	
	Neutral	0	1	55	11.58%	98.21%	20.72%	
IPC	Positive	80	5	110	90.91%	41.03%	56.54%	59.77%
	Negative	8	171	82	96.61%	65.52%	78.08%	
	Neutral	0	1	55	22.27%	98.21%	36.30%	
SWNC	Positive	149	38	8	64.22%	76.41%	69.79%	71.68%
	Negative	76	176	9	79.64%	67.43%	73.03%	
	Neutral	7	7	42	71.19%	75.00%	73.04%	

Table 13
Dataset 6 experiment and results.

Dataset 6		Confusion matrices			Results			
		Positive	Negative	Neutral	Precision	Recall	F-measure	Accuracy
Proposed	Positive	443	45	4	88.60%	90.04%	89.31%	85.90%
	Negative	45	366	6	83.18%	87.77%	85.41%	
	Neutral	12	29	50	83.33%	54.95%	66.23%	
EEC	Positive	38	4	450	63.33%	7.72%	13.77%	13.40%
	Negative	20	11	386	57.89%	2.64%	5.05%	
	Neutral	2	4	85	9.23%	93.41%	16.80%	
IPC	Positive	209	54	29	74.64%	42.48%	54.15%	52.90%
	Negative	62	243	112	80.46%	58.27%	67.59%	
	Neutral	9	5	77	18.42%	84.62%	30.26%	
SWNC	Positive	315	172	5	84.45%	64.02%	72.83%	74.20%
	Negative	35	376	6	66.55%	90.17%	76.58%	
	Neutral	23	17	51	82.26%	56.04%	66.67%	

Let *PE* denote a set of positive emoticons

$$\mathrm{PE} = \{\text{Set of Positive Emoticons}\}.$$

Let *NE* denote a set of negative emoticons

$$\mathrm{NE} = \{\text{Set of Negative Emoticons}\}.$$

The emoticon score S_e is calculated as:

$$\mathrm{Score}(e) = \begin{cases} 1, (w_x \in W) \wedge (t \in T) \wedge (w_x \in \mathrm{PE}) \\ -1, \left(w_y \in W\right) \wedge (t \in T) \wedge \left(w_y \in \mathrm{NE}\right) \\ 0, (w_z \in W) \wedge (t \in T) \wedge (w_z \notin \mathrm{PE}) \wedge (w_z \notin \mathrm{NE}) \end{cases}$$

where w_x, w_y and w_z are words belonging to set of words W and t is a tweet from the set of tweets T.

3.3.2. IPC score calculation

IPC uses 'bag of words' approach. Words are domain independent. Each word in the list has been classified as positive/negative. We have to provide words in correct spelling to be classified by IPC. Every word has the same weight. There may be a combination of positive/negative words in a tweet which may result in incorrect classification of tweet as neutral. There may be unrecognized words in the tweet resulting in incorrectly classifying the tweet as neutral. These problems can be solved by using SWNC.

A word is identified by splitting the refined tweet using the word separators like space, comma, semi-colon and full stop. The IPC is an improvement over the technique proposed in [16] by using a richer trained data set and pre-processed text. The set of positive and negative words are created from the Bing Liu list [21] and the Bill McDonald list [22]. The word count in the Bing Liu list is: 2006 positive words and 4784 negative words which makes 6790 in total whereas the word count in the Bill McDonald list is: 354 positive words and 2349 negative words which makes 2703 words in total. The total trained word count for the IPC is 9493. The sample sets of positive and negative words are given in Table 3 and Table 4. If the word is found in the positive set then it is declared as positive. It is declared negative if found in the negative set. If the word is not found in both the sets, we declare it as neutral. Total positive and negative words are counted and the sum is calculated. A score of 1 is assigned to the refined tweet, if the sum is greater than zero. A score of −1 is assigned if the sum is less than zero. A score of zero indicates that the calculated sum is zero.

Let *PW* be a set of positive words

$$\mathrm{PW} = \{\text{Set of Positive Words}\}.$$

Let *NW* be a set of negative words

$$\mathrm{NW} = \{\text{Set of Negative Words}\}.$$

The list of words score S_w is calculated as:

$$\mathrm{Score}(w) = \begin{cases} 1, (w_x \in W) \wedge (t \in T) \wedge (w_x \in \mathrm{PW}) \\ -1, \left(w_y \in W\right) \wedge (t \in T) \wedge \left(w_y \in \mathrm{NW}\right) \\ 0, (w_z \in W) \wedge (t \in T) \wedge (w_z \notin \mathrm{PW}) \wedge (w_z \notin \mathrm{NW}) \end{cases}$$

where w_x, w_y and w_z are words belonging to set of words W and t is a tweet from the set of tweets T.

3.3.3. SWNC score calculation

SWNC assigns different sentiment weights to different words. It also depends on the how the word is being used in the sentence i.e. identification of 'part of speech' for the word is necessary to be classified by SWNC.

Similar to the previous step, a word is identified by splitting the refined tweet using the word separators like space, comma, semi-colon and full stop. The sentiment value of each word is calculated by calling the SentiWordNet. Sentiment weight for each word is found and the sum is calculated by adding each of the sentiment weights. A score of 1 is assigned to the refined tweet, if the calculated sum is greater than zero. A score of −1 is assigned if the sum is less than zero. A score of zero indicates that the calculated sum is zero.

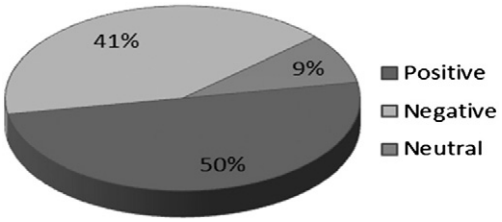


Fig. 5. Distribution of overall dataset tweets.

The SWNC score S_s is calculated as:

$$\text{Score}(s) = \begin{cases} 1, (w_x \in W) \wedge (t \in T) \wedge (\text{weight}(w_x) > 0) \\ -1, \left(w_y \in W\right) \wedge (t \in T) \wedge \left(\text{weight}\left(w_y\right) < 0\right) \\ 0, (w_z \in W) \wedge (t \in T) \wedge (\text{weight}(w_z) < 0) \end{cases}$$

where w_x, w_y and w_z are words belonging to set of words W and t is a tweet from the set of tweets T and weights are calculated using SentiWordNet dictionary.

3.4. Classifying the tweet

First we perform the EEC based classification, next IPC classification is done and lastly SWNC based classification is performed. If the result of EEC is a neutral tweet, we perform IPC and if it is still classified as neutral, we move to SWNC. If all the three techniques classify the tweet as neutral, we declare it as neutral. Otherwise, we classify it into positive or negative as declared by the classifiers. This helps to reduce the number of neutral tweets which was a major issue in previous techniques. The results also indicate that this classification procedure is more accurate than its predecessors.

The final classification is done using the three scores as below:

$$\text{Class} = \begin{cases} \text{Positive}, (S_e > 0) \vee (S_e = 0 \wedge S_w > 0) \vee (S_e = 0 \wedge S_w = 0 \wedge S_s > 0) \\ \text{Negative}, (S_e < 0) \vee (S_e = 0 \wedge S_w < 0) \vee (S_e = 0 \wedge S_w = 0 \wedge S_s < 0) \\ \text{Neutral}, (S_e = 0) \wedge (S_w = 0) \wedge (S_w = 0) \end{cases}$$

where S_e, S_w and S_s are scores from EEC, IPC and SWNC respectively.

4. Results and discussion

The datasets were generated using the data acquisition module. The experiments have been conducted using 6 different datasets. The experiments are performed using 2116 random tweets. These tweets were collected from Twitter using twitter streaming API. Random tweets with different search strings at different times were considered for analysis. Crowdsourcing method was used to gather human judgments as mentioned in [12]. Different sets of tweets were distributed among students, teachers, and industry people in such a way that we had at least 5 judgments for each tweet. Then majority voting was applied to classify the tweet. This process ensures that a good set of judgments are available for testing the classifier performance.

The upper limit of number of tweets provided by the API is around 100 in one attempt, so we have to query again to get more tweets as they come along as the time passes. In this way we collected 6 sets of tweets. Each set of the tweet is about a certain object, personality or event. Table 5 shows some examples of positive, negative and neutral tweets. The count of each data set is given in Table 6.

Confusion matrices, precision, recall, F-measure and accuracy are used for evaluation of the proposed framework and comparison with other techniques. The confusion matrix is defined in Table 7.

The diagonal elements (tpA, tpB, tpC) in the confusion matrix present the correctly classified data for each class whereas all other elements show incorrectly classified data.

Precision is defined by the fraction of true positives against both true positives and false positives (all positive results). Mathematically:

$$\text{PrecisionA} = \frac{\text{tpA}}{\text{tpA} + \text{eBA} + \text{eCA}}$$

where tpA is the number of true positive predictions for the class A and eBA, eCA are false positives.

Recall is the proportion between correctly classified positives by the classifier and manual classified positives (true positives + false negatives). Mathematically:

$$\text{RecallA} = \frac{\text{tpA}}{\text{tpA} + \text{eAB} + \text{eAC}}$$

where tpA is the number of true positive predictions for the class A and eAB, eAC are false negatives.

F-measure is the harmonic mean of both precision and recall. Mathematically:

$$F\text{-measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Accuracy is defined as the proportion of true positive, true negatives and true neutrals (true results) from all the given data.

$$\text{Accuracy} = \frac{\#\ \text{True Positives} + \text{True Negatives} + \text{True Neutrals}}{\#\ \text{True Positives} + \text{False Positives} + \text{True Negatives} + \text{False Negatives} + \text{True Neutrals} + \text{False Neutrals}}.$$

The classification algorithm runs on test datasets and processes each tweet. It classifies the tweets into positive, negative and neutral classes.

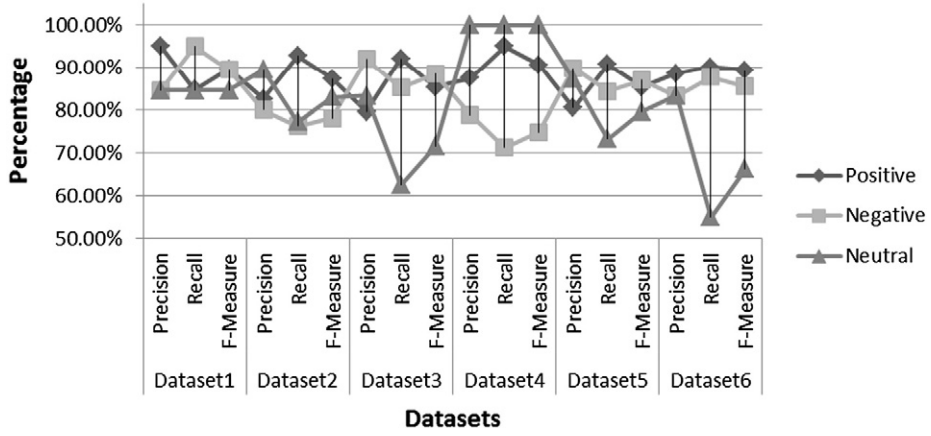


Fig. 6. Precision, recall and F-measure of proposed algorithm.

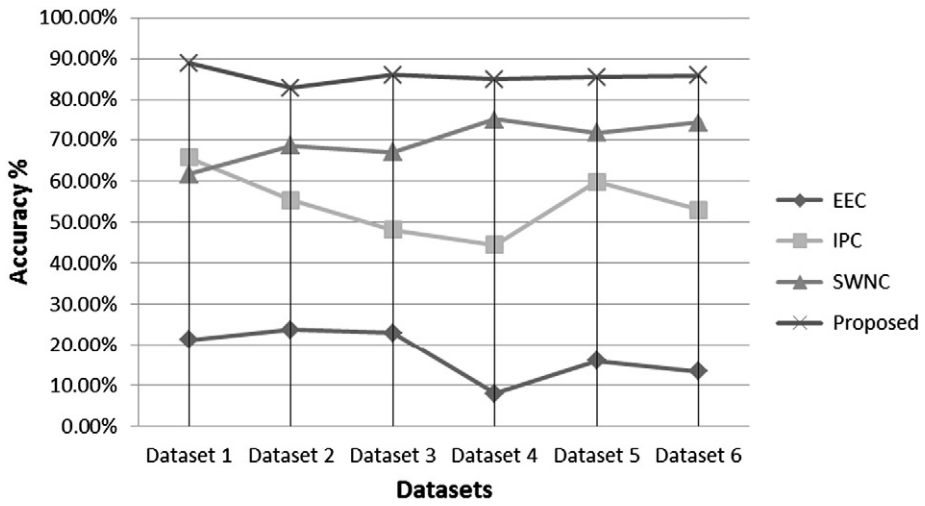


Fig. 7. Classifiers accuracy comparison using different datasets.

Confusion matrices, precision, recall, F-measure and accuracy are calculated for the proposed TOM framework and also for the EEC, IPC & SWNC. The results are used to verify the superiority of the proposed classifier. Tables 8–13 show these results for all the datasets.

The overall dataset is shown in Fig. 5. We have used a total of 2116 tweets which are classified as 1058 positive, 872 negative and 186 neutral by crowd sourcing method.

The evaluation results of precision, recall and F-measure of complete datasets are shown in Fig. 6.

We have compared the accuracy of the proposed technique with other related techniques of sentiment analysis. It is clear from the comparison that the proposed algorithm shows better accuracy for classification. The graphical representation of comparison is done in Fig. 7.

Pre-processing, EEC, IPC and SWNC play a major role in the resolution of sparsity issue. The pre-processing step is involved indirectly as it prepares the data that is worked upon by the EEC, IPC and SWNC. As a result, the proposed framework is successfully able to label the tweets without any training dataset using domain independent features like words and emoticons.

5. Conclusion and future work

The research paper has proposed a new algorithm for twitter sentiment analysis and it is based on three way classification algorithm. We have also discussed challenges that are faced during sentiment analysis and proposed the algorithm that resolves these issues and increases the classification accuracy effectively reducing the number of classified neutrals. The results of the proposed framework show great improvement when comparing with similar work. We have achieved an average accuracy of 85.7% with 85.3% precision and 82.2% recall. Future research directions include the development of a web application in order to compare the performance of our algorithm with other applications like TweetFeel & Sentiment140 and the use of supervised learning algorithms to further increase the accuracy.

References

[1] A. Cui, M. Zhang, Y. Liu, S. Ma, Emotion Tokens: Bridging the Gap among Multilingual Twitter Sentiment Analysis, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 238–249.
[2] A. Bifet, E. Frank, Sentiment Knowledge Discovery in Twitter Streaming Data, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 1–15.
[3] A. Bifet, G. Holmes, B. Pfahringer, MOA-TweetReader: real-time analysis in twitter streaming data, in: T. Elomaa, J. Hollm'en, H. Mannila (Eds.), DS 2011, LNCS 6926, Springer-Verlag, Berlin Heidelberg, 2011, pp. 46–60.
[4] S. Ye, S.F. Wu, Measuring message propagation and social influence on Twitter.com, in: L. Bolc, M. Makowski, A. Wierzbicki (Eds.), SocInfo 2010, LNCS 6430, Springer-Verlag, Berlin Heidelberg, 2010, pp. 216–231.
[5] S. Argamon, K. Bloom, A. Esuli, F. Sebastiani, Automatically determining attitude type and force for sentiment analysis, in: Z. Vetulani, H. Uszkoreit (Eds.), LTC 2007, LNAI 5603, Springer-Verlag, Berlin Heidelberg, 2009, pp. 218–231.
[6] X. Fu, Y. Guo, W. Guo, Z. Wang, et al., Aspect and sentiment extraction based on information-theoretic co-clustering, in: J. Wang, G.G. Yen, M.M. Polycarpou (Eds.), ISNN 2012, Part II, LNCS 7368, Springer-Verlag, Berlin, Heidelberg, 2012, pp. 326–335.
[7] A. Nagy, J. Stamberger, Crowd sentiment detection during disasters and crises, Proceedings of the 9th International ISCRAM Conference — Vancouver, Canada, 2012.
[8] A. Montejo-Raez, E. Martınez-Camara, M.T. Martın-Valdivia, L.A. Urena-Lopez, RandomWalk weighting over SentiWordNet for sentiment polarity detection on Twitter, Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, 2012, pp. 3–10.
[9] R. Ortega, A. Fonseca, M. Mendoza, Y. Guti'errez, SSA-UO: unsupervised Twitter sentiment analysis, in: A. Montoyo (Ed.), Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, pp. 501–507, (Atlanta, Georgia).
[10] F. Bravo-Marquez, M. Mendoza, B. Poblete, Combining Strengths, Emotions and Polarities for Boosting Twitter Sentiment Analysis, WISDOM'13, Chicago, IL, USA, 2013.
[11] J. Kim, J. Yoo, H. Lim, H. Qiu, Z. Kozareva, A. Galstyan, Sentiment Prediction using Collaborative Filtering, Association for the Advancement of Artificial Intelligence, 2013.
[12] R. Machedon, W. Rand, Y. Joshi, Automatic Classification of Social Media Messaging using Multi-Dimensional Sentiment Analysis and Crowdsourcing, 2013, 2013. http://dx.doi.org/10.2139/ssrn.2244353 (Available at SSRN: http://ssrn.com/abstract=2244353).
[13] A. Balahur, Sentiment Analysis in Social Media Texts, 2013, pp. 120–128, Atlanta, Georgia.
[14] J. Read, Using emoticons to reduce dependency in machine learning techniques for sentiment classification, Proceedings of the ACL Student Research Workshop, 2005, pp. 43–48.
[15] S. Baccianella, A. Esuli, F. Sebastiani, SENTIWORDNET 3.0: an enhanced lexical resource for sentiment analysis and opinion mining, http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf(Accessed 4 Feb 2013).
[16] B. Liu, S. Li, W.S. Lee, P.S. Yu, Text classification by labeling words, Proceedings of the National Conference on Artificial Intelligence, AAAI Press; MIT Press, Menlo Park, CA; Cambridge, MA; London, 2004, pp. 425–430.
[17] L. Barbosa, J. Feng, Robust sentiment detection on twitter from biased and noisy data, Proceedings of COLING, 2010, pp. 36–44.
[18] A. Go, R. Bhayani, L. Huang, Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford, 2009.
[19] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, Proceedings of LREC, 2010.
[20] Twitter4J, http://twitter4j.org/en/index.html, Accessed 4 Feb 2013.
[21] Bing Liu list of words, http://www.cs.uic.edu/ liub/FBS/opinion-lexicon-English.rar, Accessed 16 Feb 2013.
[22] Bill McDonald list of words, http://www3.nd.edu/mcdonald/Word_Lists.html, Accessed 16 Feb 2013.

Saba Bashir is an Assistant Professor in Computer Science Department at Federal Urdu University of Arts, Science & Technology, Pakistan. She is also a PhD research scholar at NUST, Pakistan. Her research interest lies in predictive systems, web services and object oriented computing. She has published more than 8 research papers in international conferences & journals.

Farhan Hassan Khan has been working as a Project Manager in a software development organisation in Pakistan since 2005. He is also a PhD research scholar at NUST, Pakistan. His research interest lies in text mining, web service computing and VoIP billing products. He has published many research papers in international conferences & journals.

Dr. Usman Qamar is an Assistant Professor in Computer Engineering Department at CE&ME, National University of Science & Technology, Pakistan. He completed his PhD (Information Systems) from the University of Manchester, UK. His research interest lies in Data Mining, Outlier Detection, and Feature Selection.