IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Emotion Detection and Recognition from Text using Machine Learning

Shaikh Abdul Salam[1*], Rajkumar Gupta[2]

[1] CSE, RITM, Maharshi Dayanand University, Palwal, India
[2] CSE, RITM, Maharshi Dayanand University, Palwal, India

[*]*Corresponding Author: salamshaikh9@gmail.com, Tel.: +91-98219-67863*

Available online at: www.ijcseonline.org

Accepted: 12/Jun/2018, Published: 30/Jun/2018

Abstract— In today's technological world, a majority of users across the world have access to Internet for communication via text, image, audio and video. People from diverse backgrounds exchange information on current scenarios and project their own views on them over social media. There is a need to understand and recognize the behavior of such large text information on people by analyzing their emotions. The paper focuses on data obtained from one of the most popular social media - Twitter by analyzing live as well as past feeds and getting emotions from them. The twitter data required in English language is converted into a vector of eight emotions and supervised learning techniques such as K-means, Naive Bayes and SVM is used to determine label identifying one of the basic emotion family. At the end, a comparative study of the performance of different classifiers is discussed.

Keywords— sentiment, machine learning, emotion detection, twitter, SVM, Naive Bayes.

I. INTRODUCTION

Social media has become an integral part of the people in 21st century. Due to rapid progress in Information & Technology sector, people have access to any kind of information at the click of a button. Moreover, with the invent of smart-phones and 4G networks, even people from the remote areas are getting connected to Tier 1 and Tier 2 cities. With the growing population in countries like India, it has led to tremendous growth in the number of people using social networks.

Social networks like Facebook, WhatsApp, Twitter, etc. has eliminated the gap between lives of people. One of the reasons to use these social networks to know the current happenings around them and to express their views and suggestions in the form of likes, share, tweets, polls, email, etc. This has created a new category of people called netizens. Communication via social media is done in the form of text, image, audio and video which contains information and consumes space, memory and Internet bandwidth. All these activities done on social media has resulted into vast amount of information being generated on a daily basis. Social media analysis has become a interesting field of research to understand the behavior and thoughts of people in response to social, economic, cultural, educational and all others activities happening around the world [1].

Social media data is in the form of unstructured data since people are from diverse backgrounds with different race, culture, language and standard of living project their ideas, views, opinions, expressions and so on the Internet. So, it has become a challenge in recent years to extract valuable information from these ever-growing data in the form of posts, emails, blogs, micro-blogs, tweets, reviews, comments, polls and surveys on the Web about an individual, an organization or government in the process of decision-making [1]. These opinionated data not only exist on the Web but also within the large organizations like Google, Microsoft, Hewlett-Packard, SAP and SAS to know the opinions of their employees spread across the continents in the form of customer feedback collected from emails and call centers or results from surveys conducted by organizations. This has created strong interest for research in sentiment analysis.

Sentiment Analysis has emerged as a popular solution to gain deeper insights such as knowing how happy our people are, predicting box-office revenues from movie reviews, predicting election results and stock market from Twitter data, gaining feedback before releasing a new product and visualizing of customers feel about product or service run by multinational companies. From all these sources, Twitter has become one of the prominent leads in gathering and expressing users opinions publicly within a short message. With 140-character limit, people have freedom to express their feelings, opinions, appraisal, emotions, attitudes, evaluations and sentiments on present or past events, trends, lifestyles, government, health care, education, business, politics and all kinds of activities around the world. Twitter statistics report states that on every second on an average around 6,000 tweets are tweeted on Twitter which results to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year [2].

In this paper, emotion mining on data over Twitter is analyzed by using Machine Learning techniques on recent trends, topics, discussions over the Web to determine the effect of the textual information on the mentality of the people by identifying their emotions. This paper is organized in following sections. Related work is discussed in Section II. The proposed system details are provided in Section III followed by experiments and results in Section IV. The conclusion of this paper is provided in Section V with future work direction

II. Related Work

Work on Emotion classification has been carried out for the last three decades, emotion mining is relatively new and has gained interest among researchers due to popularity of social media platforms which has created a new dimension for freedom of expression. Researchers have proposed different methods for sentiment analysis such as lexicon-based, machine learning-based or a combination of both. Some of them are explained below:

Emotions have been classified into six basic categories by Paul Ekman (1992) [3]. According to him, these basic emotion families are distinct and are basic in the sense that other complex emotions can be derived from two or more basic emotion. The six basic emotions identified by Ekman are Anger, Sadness, Disgust, Fear, Enjoyment and Surprise. Interestingly, majority of these emotions are skewed towards negative emotions.

Bing Liu [1] identified various challenges in sentiment analysis in his publication. The first problem was to determine whether a review is subjective or objective. Subjective reviews consist of explicit emotions while objective reviews consist of implicit emotions. It is easier to work on subjective reviews as compared to objective reviews because the latter contained implicit emotions or no emotions at all. Also, he mentioned, various Bi-grams containing adjectives and adverbs are good indicators of explicit emotions and opinions which can be extracted with the help Parts of Speech (POS) tagger. Another problem was to identify the orientation of sentiment which were found to be changed when used in association with some other word e.g. ’not feeling good’ or ’rarely active’.

Anees and Jamil [4] presented a paper to find the depression level of a person by studying and finding emotions from text using emotion theories, machine learning and NLP techniques from social media data. They used binary and multi-class sentiment classification techniques at sentence level analysis and implemented using three classifiers, SVM, Naive Bayes and Maximum Entropy. They concluded by making comparison among SVM, NB and ME classifiers by adopting voting model and feature selection technique. Twitter and 20newsgroups were used as datasets and experiment results indicated that SVM displayed better results than NB and ME classifiers. It was observed that accuracy of SVM is 91%, Naive Bayes is 83% and ME is 80%.

Saima and Aman [5] addresses the task of identifying the emotions in a text by using annotated corpus. They used knowledge-based technique that classifies emotional and non-e motional sentences automatically. Two commonly used lexical resources the General Inquirer and WordNet-Affect were used to recognize emotion words present in the sentence. They used supervised machine learning techniques for classification and found SVM technique with highest accuracy 73.89% as compared to Naive Bayes with 72.08%.

The concept of contextual valence shifters presented by Livia Polanyi and Annie Zaenen [6] plays an important role while performing sentiment analysis. They describe how the basic valence of individual lexical items may have positive or negative effect or changed by context provided by other lexical items, the genre type and discourse structure of the text and cultural factors. They argue strongly that determining the author attitude must be based on a finer grained analysis of the text on all levels.

Barbara Plank and Dirk Hovy [7] suggest that social media data can provide sufficient linguistic evidence to generate large number of self-identified training tweets for gathering proper training data. They used 1.5M tweets from 1,500 different users and for each user, they downloaded their most recent tweets with minimum 100 tweets and maximum 2000 tweets to generate the final corpus in one-week duration.

III. PROPOSED SYSTEM

The Proposed system will detect 8 basic emotions as proposed by Plutchik, in the live twitter stream and classify tweets as ’anticipation’, ’enjoyment’, ’sad’, ’disgust’, ’anger’, ’surprise’, ’fear’, ’trust’ or No emotion. This classification is implemented by using supervised machine learning: - Naive Bayes and Support Vector Machines (SVM) and unsupervised machine learning: - K-Means. A comparative study of the efficiency of these three classification methods has been performed.

There are total 8 Bag of Words (BOW) created that represents each of the emotion family. The BOW is mutually exclusive i.e. word from one bag cannot be present in another bag.

$$\{Anger\} \cup \{Fear\} \cup \{Sad\} \cup \{Surprise\} \cup \{Trust\} \cup \{Disgust\} \cup \{Anticipation\} \cup \{Enjoyment\} = \varphi$$

A. Collecting Data

The Tweet Retrieval module works on multiple languages by fetching the tweets based on the selected region. So that every non-English tweet is first converted into English sentences with the help of online translators such as Microsoft Bing, Google Cloud Translate, etc. Then the converted tweet is provided to the Emotion Tagger module that converts each incoming tweet into a vector of 8 basic emotions. The reason behind using translator is based on the argument that in any text, we just need the English word that safely fall in of the 8 basic emotion categories related bag of words (BOW). These vectors are then fed into supervised learning techniques SVM and Naive Bayes which returns a emotion label for the given tweet. K-Means works in a different fashion as it needs entire dataset to work on. The result of K-Means is K clusters, where these K-Clusters are assigned a label based on the frequency distribution of emotions in the cluster.

B. Generating dataset

To generate training data set, a total of 100 words from each of the bags is used. These words are first converted into hashtags and then fed into the Tweet Retrieval System (TRS). The TRS keeps on listening to the live tweet stream and captures any tweets with the hashtag that is passed to it. Labelling of training dataset is automated when a tweet is retrieved it is assigned a label which is one of the basic emotion family to which the hashtag word belongs.

TABLE I
SAMPLE TWEETS WITH LABEL

Tweets	Tag	Label
Tweet1	#ANXIOUS	Fear
Tweet2	#AMAZE	Surprise
Tweet3	#DEPRESS	Sad

The TRS module can be used retrieve tweets based on GeoLocation, specific language, hashtags and any key word.

1) Pre-processing Data: The tweets fetched by the TRS module contains lots of garbage value which create problems while classification. These values or connectors have to be eliminated from the training data before it is passed to the next module i.e. Emotion Tagger in order to get mapped to the corresponding vector representing basic emotion. The data pre-processing involves replacing the HTTP links by URL, eliminating new lines, tabs and hash characters and converting

TABLE II
SAMPLE TWEETS WITH LABEL

Tweets	Tag	Label
URL	https://t.co.e/eJKsavbNYaoL	https?:[\r \n]
\n	so much love and care	[\n]
\t	Generally \t It is..	[\t]

2) Collecting Corpus: Total of 6000+ Tweets were collected for training purpose. These tweets are labelled and are associated with vector corresponding to one of the basic emotion family. Other meta data associated with the training set are as follows:

TABLE III
META-DATA OF SAMPLE TWEETS

ID	TIME	LAN	ANNOT	VECTOR
1223	Mon Dec 14 20:01:50	en	enjoyment	[0 0 0 0 1 0 1 0]
4533	Mon Dec 14 17:13:38	en	disgust	[0 0 1 0 1 0 0 0]
6712	Fri Nov 27 16:14:40	en	surprise	[0 0 0 0 0 0 1 0]

C. METHODOLOGY

The core component of the system is the Emotion Tagger (ET) that takes a sentence as input, performs calculation and gives the vector as output. These vectors are then used to train our supervised learning classifiers - SVM and Naive Bayes. Linear Kernel for SVM is used for implementation. After the SVM and Naive Bayes classifiers are trained, live tweets can be fed into them to determine the emotion type.

TABLE IV
SAMPLE TWEETS WITH VECTOR

TEXT	VECTOR
I wonder if Chelsea can cause an upset and beat Leicester.	[0 0 0 0 1 0 1 0]
In my opinion the worst side effect from pills is nausea!!	[0 0 1 0 1 0 0 0]
You seem to amaze me. Smh. You aren’t showing anything.	[0 0 0 0 0 0 1 0]

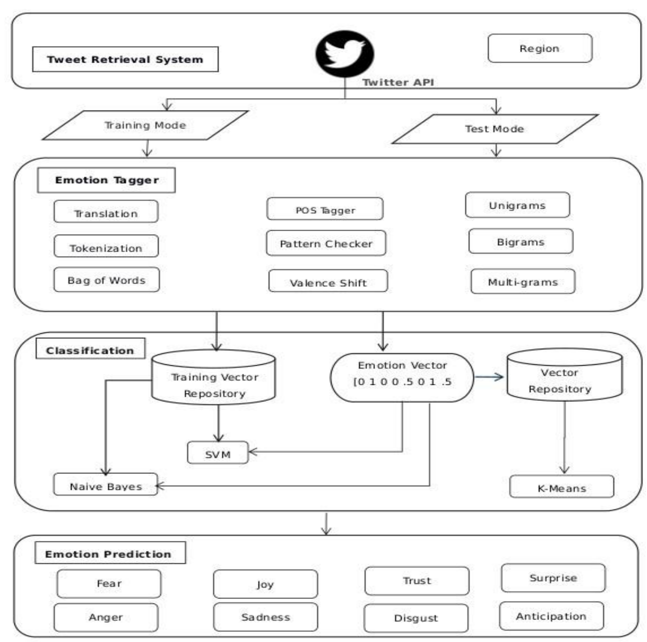

Fig. 1. Proposed System.

1) Supervised Learning: The Emotion Tagger module which converts each incoming tweet into a vector of emotion works at the syntactic label of a sentence. Therefore, it can detect emotions subjective tweets or reviews because emotions are explicitly mentioned in them. However, it fails to recognize implicit emotions which may or may not be present in objective reviews or tweets. As mentioned earlier, a total of 8 mutually exclusive Bag of Words (BOW) is created. Each bag consists of five words expressing their corresponding emotion. These words are most frequently used to describe or express emotion, handpicked EMO words from 'General Inquirer'.

The Emotion Tagger module can detect contextual valance shift which occurs under certain conditions. They are described below:

• It can detect valance shift due to presence of negation like NOT.

• It can detect intensifiers like very, so deeply.

• It can detect presupposition items like barely, even.

• It can detect words like rarely, seldom which decrease the intensity of emotion.

• Certain words like" neither nor" which nullifies intensity of emotion. e.g." I'm neither happy nor sad" can be detected by tagger.

• It can detect uni-gram, bi-gram and multi-grams.

Emotion Tagger module handles all these situations differently and assigns intensity to each emotion according to following rules.

TABLE V
INTENSITY OF CONTEXTUAL VALANCE SHIFT

Context	Example	Intensity
Negation	NOT	(-)ve
Intensifiers	Very	1.5
Presupposition	barely	0.5
Nullifiers	neither nor	0

Whenever the ET module encounters a negation which is associated with emotion word, it changes the emotion family to its opposite emotion family. e.g." I'm not happy (Enjoyment family)" is considered as" I'm sad (sad family)". In order to detect bi-gram and multi-grams, ET tags each sentence with the help of NLTK POS tagger. These patterns can be present in following forms.

TABLE VI
NLTK POS TAGGER

Word Extent	Example	POS Pattern
Bi-gram	rarely sad	RB JJ
Multi-gram	inability to perform good	NN TO VB JJ

The overall working of the ET module is described in the following diagram

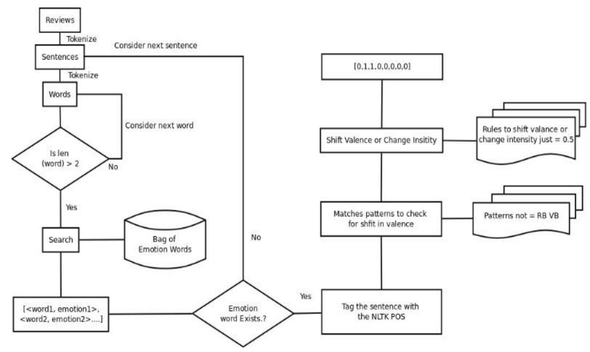

Fig. 2. Emotion Parser in Supervised learning.

2) Unsupervised Learning: K-Means is a clustering algorithm but here it is used as classification. It is unsupervised learning algorithm so it doesn't require any training data set. When K-Means algorithm is implemented on the data set collected for training, the output is K-clusters. In order to assign a label to these clusters, the frequency distribution of emotion of each cluster is calculated. The emotion family whose frequency is highest is then assigned as the label to that cluster. So, the label of a tweet is same as the label of a cluster to which it belongs. The entire labelling process is repeated nTimes, in our case, it is 17. This is to mitigate the effect of random initialization done in K-Means.
The entire process is described in the following diagram.

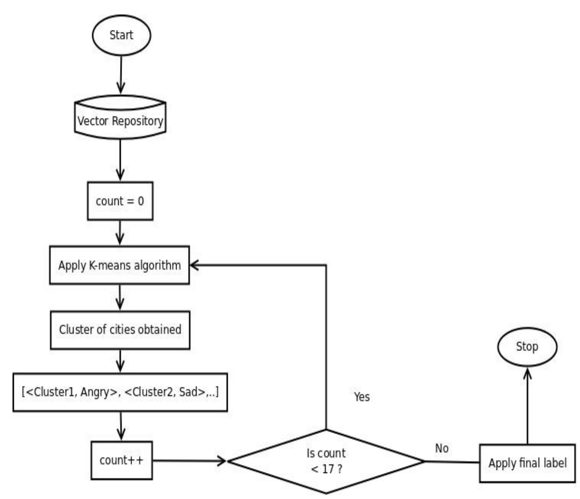

Fig. 3. Emotion Parser in Unsupervised learning.

IV. EXPERIMENTS AND RESULTS

Accuracy Baseline 1: 23.45 % Accuracy Baseline 2: 39.6 %

TABLE VII
PERFORMANCE TABLE

Algorithm	Accuracy	Precision	Recall
KMeans(K=8)	40.1	41.31	40.40
KMeans(K=8+1)	40.00	41.00	40.0

NBC(5-F)	64	58	64
NBC(10-F)	64	58	54
SVC(5-F)	66	56	66
SVC(10-F)	65	56	56

Accuracy Baseline 1 mentioned above has been calculated from training data set where Anger was the most prevalent emotion. For training, a total of 13000+ tweets were used.

Anger: 3220 / 13731 = 23.45%

Accuracy for Baseline 2 above has been referred from the work of Soumaya Chaffar and Diana Inkpen[9].

TABLE VIII
CLASSIFICATION REPORT OF LINEAR SVC(5-FOLD)

Emotion	precision	recall	f1-score	support
anger	0.80	0.91	0.85	254
anticipation	0.00	0.00	0.00	79
disgust	0.63	0.84	0.72	73
enjoyment	0.60	0.86	0.70	171
fear	0.61	0.79	0.69	173
sad	0.56	0.61	0.64	199
surprise	0.56	0.51	0.53	83
trust	0.00	0.00	0.00	87

It is observed that SVC approach (with 5-Fold validation) is the best performer. It is evident from the above classification report that NBC and SVM works better that K-Means. Also 5-Fold cross validation performed better than 10-Fold. And K-Means performance doesn't change with K=9 and K=10.

V. CONCLUSION

The most important part is getting good quality of data. Often tweets retrieved consists of only URLs and hence pre-processing data still remains one of the most crucial steps which needs to be improved. Secondly, only explicit emotions are explored in the tweet so detecting implicit emotions in objective sentences can in included in future work. At present, each sentence is treated as individual unit. Further the association between the sentences can be included. The proposed system works at the syntactic level only, semantic level parsing can be implemented further. Finally, the Bag of words needs to be more exhaustive and hence more words describing emotions can be added to these bags.

REFERENCES

[1] Bing Liu (2009) Sentiment Analysis and Subjectivity, Handbook of Natural Language Processing, Second Edition (editors: N. Indurkhya and FJ Damerau)

[2] Santhoshi Kumari, Narendra Babu, Real Time Analysis of Social Media Data to Understand People Emotions Towards National Parties, 8th ICCCNT 2017, July 3 - 5, 2017, IIT Delhi, Delhi.

[3] Ekman, P (1992), An Argument for Basic emotions, Cognition and Emotion, 6(3-4), (Pg. - 169-200).

[4] Sentiment Analysis of Social Networking Sites Data using Machine Learning Approach for the Measurement of Depression, Anees Ul Hassan, Jamil Hussain, Sungyoung Lee, ICTC 2017, (Pg. - 138-140)

[5] Saima Aman and Stan Szpakowicz (2007), Identifying Expressions of Emotion in Text, V. Matouek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 196205, 2007. Springer-Verlag Berlin Heidelberg 2007

[6] Livia Polanyi and Annie Zaenen (2006), Contextual Valence Shifters, Computing attitude and affect in text: Theory and applications, (Pg. 1-10)

[7] Plank B, Hovy D (2015), Personality Traits on Twitter-or How to get 1,500 Personality Tests in a Week, In 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015) (Pg. 92)

[8] Lionel Martin and Pearl Pu (2014), Prediction of Helpful Reviews Using Emotions Extraction. AAAI Conference on Artificial Intelligence Twenty-Eight AAAI Conference on Artificial Intelligence.

[9] Soumaya Chaffar and Diana Inkpen (2011) Using a Heterogeneous Dataset for Emotion Analysis in Text, Proceeding Canadian AI11 Proceedings of the 24th Canadian Conference on Advances in Artificial Intelligence (Pg. 62-76).

[10] Emotion and Polarity Prediction from Twitter, Rebeen A Hamad, Saeed Alqahtani Mercedes Torres, Computing Conference 2017, 18-20 July 2017, London UK, (Pg. 297-306)

[11] Exploring Human emotion via Twitter, Abu Z Riyadh, Nasif Alvi, Kamrul Hasan, 2017 20th International Conference on Computer and Information Technology (ICCIT), 22-24 December 2017.

[12] Emotion Detection from the SMS of the Internet, Uma N, Priyanka K, Aditi M, Dr. Dhananjay K., 2013 IEEE Recent Advances in Intelligent Computational Systems (RAICS). (Pg. 316 - 321)

[13] Twitter Sentiment Analysis and Opinion Mining, Ravikiran Janardhana, Dept. of CSE, University of North Carolina.

[14] Framework for Emotion Detection and Classification of English and Non-English Tweets, Anup Prasad and Mayank Jaglan, Indiana University Bloomington, School of Informatics and Computing.

Authors Profile

Mr. Shaikh Abdul Salam completed his Bachelors of Engineering from University of Mumbai, India in 2012. He is currently pursuing M.Tech (CSE) in Rattan Institute of Technology & Management and currently working as Assistant Professor in Department of Computer Engineering, University of Mumbai, India. He is a life member of ISTE. His main research work focuses on NLP, Machine Learning, Web Technologies, Network Security and Database systems. He has 5 years of teaching experience and 1 year of Industry Experience.

Mr. Raj Kumar pursed B.E (CSE) from Maharshi Dayanand University, Rohtak (HR), India in year 2007 and M.Tech (CSE) from Maharshi Dayanand University, Rohtak (HR), India in year 2013. He is currently currently working as Assistant Professor in Department of Computer Science & Engineering, Rattan Institute of Technology & Management, Maharshi Dayanand University, India since 2013. His main research work focuses on Cryptography Algorithms, Network Security, Cloud Security and Privacy, Big Data Analytics, Data Mining, IoT and Computational Intelligence based education. He has 5 years of teaching experience.

