Mood Analysis in Twitter Using Machine Learning Techniques

Sharad Kakran
Department of Computer Science & Engineering
SRM University
Kattankulathur, Chennai
sharadkakran@gmail.com

Shantanu Garg
Department of Computer Science & Engineering
SRM University
Kattankulathur, Chennai
shantanu.garg@ymail.com

S Thenmalar
Department of Computer Science & Engineering
SRM University
Kattankulathur, Chennai
thenmalar.s@ktr.srmuniv.ac.in

***Abstract*—**With a rapid increase in the outreach of the internet today, people have taken the internet as a means to express their opinion on topics ranging from their preferred political party to their favourite restaurant. Social media websites such as twitter are one such place where people can express themselves. Twitter generates a vast amount of sentiment rich data which can be studied in order to identify the effects of sentiment classification. This information can be used to understand the perception of twitter users on various subjects. In this paper, we try to use machine learning techniques such as LSA, NB, SVM, KNN to analyse twitter posts on various subjects. Mood mining deals with identifying and extracting the sentiment content of a text unit using NLP, Statistics or machine learning techniques. We present a new feature for classifying the tweets in five different moods - happy, sad, fun, love and hate.

***Index Terms*—**Latent Semantic Analysis (LSA), Naive Bayes Algorithm (NB), Support Vector Machine (SVM), K-Nearest Neighbours Algorithm (KNN), Classification, Sentiment Analysis, Twitter

I. INTRODUCTION

Mood analysis is an emerging field of research. Today, where we have trillions of gigabytes of worth of data on the internet in the form of blogs, status updates, tweets etc., sentiment analysis can help us in establishing and further understanding the pattern between this data. Generally, mood mining involves classifying the presence of positive and negative emotions in a text document. In this model, our approach goes beyond previous work in that our model further tries to classify the mood of the author of the text at the time of the writing. Segregating twitter posts according to the various moods can help us in predicting the outcomes of various events like the performance of stock market or the result of elections. Also, this methodology can be used by companies and corporations in an attempt to better understand their public perception helping them save a lot of money by allowing them to focus on their marketing and communication effects accordingly [1][2].

Mood analysis can be conducted on various levels ranging from parse level to fine level. In the scope of this paper, we use sentence level mood mining which falls somewhere in between the parsing level and fine level [3]. Unlike the previous work done in this field, we compare the performance of machine learning algorithms on two types of twitter posts i.e. context-specific and random tweets.

II. RELATED WORK

There are many different techniques and methods presented in the literature by many researchers for the sentimental analysis of twitter posts. There has been a vast amount of research conducted in the area of sentiment classification. There are many different techniques and methods presented in the literature by many researchers for the sentimental analysis of twitter posts. There has been a vast amount of research conducted in the field of sentiment classification. In general, most of the research has been done in classifying larger pieces of texts such as movie reviews or product reviews [4]. Twitter is slightly different from reviews as the maximum characters allowed per post is 140. Also, unlike reviews tweets are more casual in nature and are less thoughtfully composed. Yet, twitter is an excellent medium for companies to collect feedback to study consumer behaviour. Previously researchers have built sentiment classifiers that are able to determine positive, negative and neutral sentiment for a sentence [5]. In a similar research, Divya et al used lexicons and Naïve Bayes classifiers in order to implement mood classification in twitter data [6].

Barbosa et al in their research designed a two-step automatic sentiment analysis method for the purpose of classification of tweets. For the purpose of the research they used a noisy data set. They classified the tweets into two categories: - subjective and objective following that they further classified the subjective tweets into positive and negative [7].

Johan et al in their research paper modelled public mood and emotion using sentiment analysis on twitter. They used a psychometric aggregator to extract the six mood states from the twitter data. Next they computed a six dimensional vector for each day in their timeline. They observed that events in the social, cultural and economic sphere can have a huge impact on the public mood [8].

Mike Thelwall and Kevan Buckley in their research paper used two new approaches, lexicon extension and mood setting in order to improve the accuracy of their algorithm of topic-specific lexical sentiment strength detection for the social web. It was observed that both the methods were able to improve sentiment analysis performance on the corpora when the topic focus is tightest [9].

III. PROPOSED SOLUTION

The proposed system architecture for the mood classifier is shown in Figure 1. There are five main steps in the process and they are as follows:

A.) Creation of Dataset
B.) Segregation of Dataset in two categories i.e. context-specific tweets and random tweets
C.) Pre-Processing the data
D.) Modelling
E.) Evaluating the accuracy of each algorithm

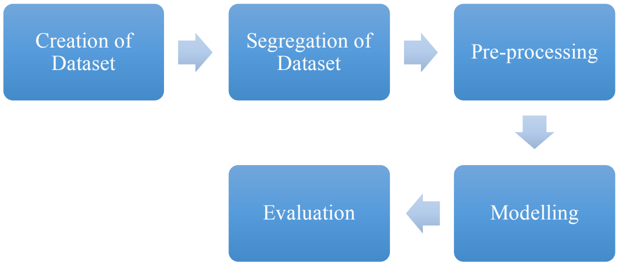

Figure 1: System Architecture Block Diagram

A.) Creation of Dataset

For the purpose of this research, we create our own dataset. We download tweets on various topics using the API's provided by Twitter.

B.) Segregation of Dataset in two categories i.e. context-specific tweets and random tweets

Next, we segregate the tweets into two categories for a better understanding of the importance of context in classification. We divide the tweets into context-specific tweets and random tweets. Context-specific tweets are the tweets relating to a particular topic. In this paper, we use the tweets on the topic of Mother's day. Random tweets are a collection of tweets ranging from a wide range of topics.

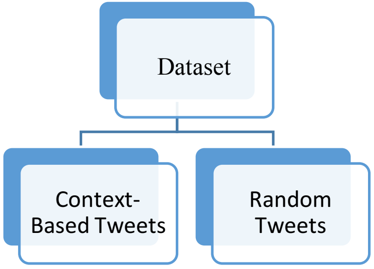

Figure 2: Segregation of Tweets

C.) Pre-Processing the data

After we have segregated the data into different categories, we perform a pre-processing step.
In this step we perform four functions that are as follows:

1.) Data Cleaning
2.) Data Splitting
3.) Feature Extraction
4.) Feature Selection

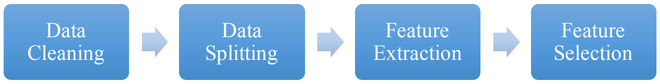

Figure 3: Pre-Processing Block Diagram

Twitter posts contain a lot of misspellings and slang words. Data cleaning is performed in order to get rid of such instances. In Data splitting the dataset is divided into training data and test data. Training data is used to train the model and test data is used to check the performance of the model.

Dataset	Fun	Happy	Hate	Love	Sad
Training	655	1042	465	1346	1058
Test	281	447	199	577	454

Table 1: Statistics of the dataset used for Random tweets

Dataset	Happy	Love	Sad
Training	1556	953	978
Test	648	327	426

Table 2: Statistics of the dataset used for Context-specific tweets

Feature extraction involves extracting of features from the entire dataset. First, we make a bag of words. Bag of worlds involves the counting of instances of each word in a particular sentence. Next, we use POS. POS is part of Speech, here we tag each sentence to its respective part of speech. In order to increase the frequency, we also used bi-grams and tri-grams. Traditionally, researchers remove the stop words, punctuation symbols and emoticons in order to reduce the size of data. We observed that removing them resulted in a decreased accuracy. Hence, in this paper we did not remove stop words, punctuation symbols and emoticons.
In order to save our data from running into curse of dimensionality, we reduced the size of data by using a statistical technique called the Chi-square test.

D.) Modelling

After pre-processing the data, we applied different machine learning techniques on it. To understand the effect of different machine learning algorithms on accuracy, we followed two approaches.
In the first approach, features are trained using three different algorithms- KNN, SVM, NB. Here, in our research the optimum value of number of neighbours in the KNN algorithm is 20. In the second approach in order to use the context of the sentence, we used LSA. Following the use of LSA, classification is done KNN, SVM and NB algorithms on the

result.

IV. CLASSIFICATION TECHNIQUES

In this paper, we have use four different types of classification algorithms for the purpose of mood analysis of the tweets.

A.) Latent Semantic Analysis (LSA)

Latent Semantic Analysis is also known as Latent Semantic Index (LSI). It is used extensively in text-based sentiment classification. Its purpose is to analyse text documents and understand the latent meaning of those documents. The LSA algorithm follows a two-point approach in order to classify a document. The first step is called the bag of words approach. Here, the documents are represented as “bag of words”. A count matrix is prepared where the number of instances of each word is calculated. In this step, the order of words is not important. In the second step, the algorithm tries to look for patterns for words that usually occur together in a certain kind of document. Next, we try to identify the frequency of a word in each document following which we can assign weights to each word. The most popular weighting technique is called TFIDF.

$$TFIDFi,j = (Ni,j / N*,j) * \log(D / Di) \qquad (1)$$

In this formula, TFIDF is the term frequency in the document, Ni,j represents the number of times the word i appears in the document j, N*,j represents the number of total words in the document j, D is the number of documents and Di is the number of documents in which word i appears.

B.) Naïve Bayes Algorithm (NB)

Naïve Bayes is a collection of classification algorithms which are based on the Bayes theorem. The common principle that each algorithm in the Naïve Bayes family shares is that every feature being classified is independent of the value of any other feature. Naïve Bayes is a simple algorithm than can sometimes outperform more sophisticated algorithms. The conditional probability for the Naïve Bayes is defined as :

$$P(Xj\ yj) = \_mi=1\ P(xij\ yj) \qquad (2)$$

In this formula, is the feature vector defined as $X=fx_1 ,x_2 ,....x_mg$ and y_j is the class label.

C.) Support Vector Machines (SVM)

Support Vector Machines aka SVM is a machine learning problem that is widely used in classification problems. SVM’s are based on the idea of finding the right hyperplane that can divide a dataset into two classes. SVM uses the discriminative function defined as

$$g(X) = w^T \_(X) + b \qquad (3)$$

In this formula, ’X’ is the feature vector, ’w’ is the weights vector and ’b’ is the bias vector. _ () is the non-linear mapping from input space to high dimensional feature space. ’w’ and ’b’ are learned automatically on the training set.

D.) K-Nearest Neighbours (KNN)

K-Nearest Neighbours algorithm or the KNN algorithm is a non-parametric lazy learning algorithm. This algorithm stores all available cases and classifies new cases based on similarity measure. It measures the distance between a query scenario and a set of other scenarios in the dataset. KNN algorithm is measured by a distance function.

$$dist(x, y) = {}^{n}{}_{i=1}\sum \sqrt{}(x_i)^2 - (y_i)^2 \qquad (4)$$

In this formula, We can compute the distance between two scenarios using some distance function d(x,y), where x,y are scenarios composed of N features.

V. EVALUATION

In this paper we used KNN, LSA, SVM and NB algorithms in order to classify the two categories of tweets. For the first category we classified the data into five different moods - happy, sad, hate, love and fun. In the second category in order to understand the effect of context we classified the tweets into three different moods – happy, love, sad.

All these classifiers had almost similar accuracy for this feature vector. It was observed that all the classifiers performed better on the context based tweets.

In the case of random tweets, it was observed that KNN algorithm performed the best with an accuracy of 72.69%

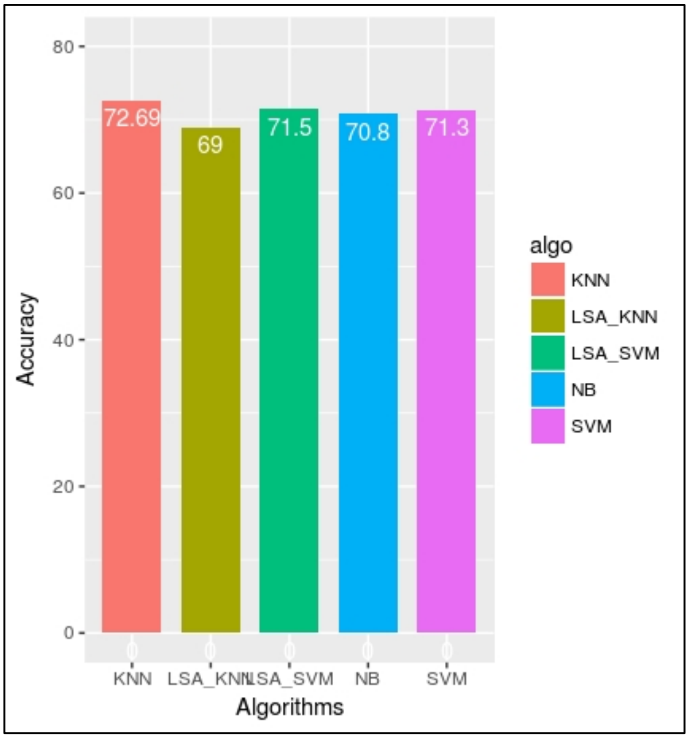

Figure 4: Performance on random tweets

In the case of context-specific tweets it was observed that the SVM algorithm performed the best with an accuracy of 93.38% followed in closely by the KNN algorithm at 93%.

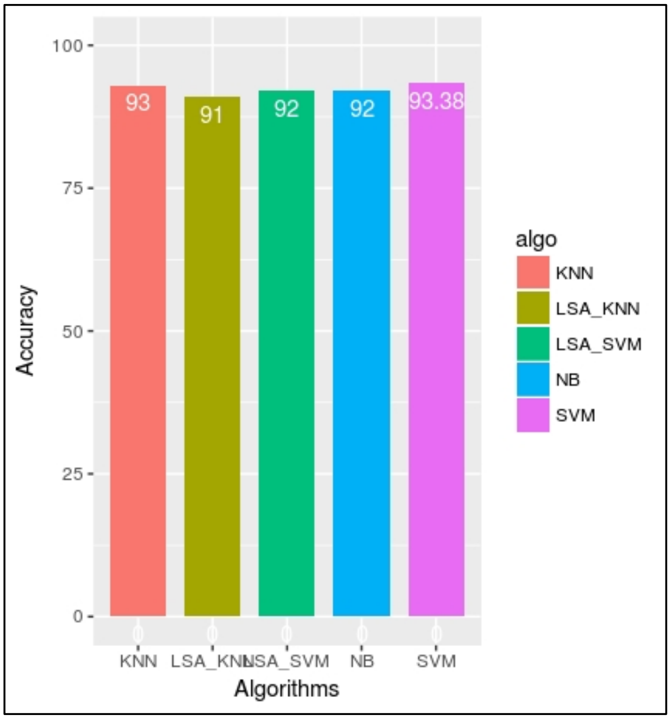

Figure 5: Performance on Context specific tweets

VI. CONCLUSION

In order to better understand the effect of context on accuracy in the proposed system we performed classification on context-specific tweets and random tweets. In both the approaches we used KNN, SVM, NB for classifying. It was observed that context-specific tweets were more accurate than random tweets. In this paper, we applied LSA algorithm to make better use of context based classification. But, due to the limited length of twitter posts, we observed that LSA algorithm did not perform as expected.

Classification accuracy of the vector is tested using different classifiers like Naïve Bayes, K- Nearest Neighbour

REFERENCES

[1] Johan Bollen, Huina Mao, Xiao Jun Zeng, “Twitter mood predicts the stock market” available *on https://arxiv.org/pdf/1010.3003&*

[2] Andranik Tumasjan, Timm. O. Sprenger, Philipp. G . Sandner, Isabelle M. welpe, “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment” available on *http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852Predicting*

[3] Yelena Mejova, “Sentiment analysis: An overview” available on *https://www.academia.edu/291678/Sentiment_Analysis_An_Overview*

[4] Bo Pang, Lillian Lee, hivakumar vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques” available on *http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf*

[5] Alexander Pak, Patrick Paroubek, “Twitter as corpus for sentiment analysis and opinion mining” available on *http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf*

[6] L. Divya Vani, D.Suneetha, “Mood classification of social media text” available on “*http://www.ijcscn.com/Documents/Volumes/vol5issue5/ijcscn2015050507.pdf*”

[7] L Barbosa, J feng, “Robust sentiment detection on twitter from biased and noisy data” available on *http://www.aclweb.org/anthology/C10-2005*

[8] Johan Bollen, Huina Mao, Alberto Pepe, “Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena” available on *http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewfile/2826/3237/*

[9] Mike Thelwall, Kevan Buckley, “Topic-Based Sentiment Analysis for the Social Web: The role of Mood and Issue-Related Words” available on *http://scitsc.wlv.ac.uk/~cm1993/papers/SentiStrength3Preprintx.pdf*