Sentiment Analysis in Twitter

Sayali P. Nazare[1], Prasad S. Nar[2], Akshay S. Phate[3], Prof. Dr. D. R. Ingle[4]

[1,2,3] Student, Dept. of Computer Engineering, Bharati Vidyapeeth College of Engineering, Maharashtra, India
[4]Professor and Head of Department, Dept. of Computer Engineering, Bharati Vidyapeeth College of Engineering, Maharashtra, India

---

Abstract - *Sentiment analysis deals with identifying and classifying opinions or sentiments expressed in source text. Microblogging today has become a very popular communication tool among Internet users. Millions of messages are appearing daily in popular web-sites that provide services for microblogging such as Twitter, Facebook. Authors of those messages write about their life, share opinions on variety of topics and discuss current issues. Because of a free format of messages and an easy accessibility of microblogging platforms, Internet users tend to shift from traditional communication tools to microblogging services. As more and more users post about products and services they use, or express their views, microblogging web-sites become valuable sources of people's opinions and sentiments. Such data can be efficiently used for marketing or social studies.*

Key Words: **Twitter API, Data set, Keyword Extraction, Classification Techniques, Reviews, Pie Chart, etc.**

1. INTRODUCTION

The age of Internet has changed the way people express their views. It is now done through blog posts, online discussion forums, product review websites etc.

When someone wants to buy a product, they will look up its reviews online before taking a decision. The amount of user generated content is too large for a normal user to analyse. So, to automate this, various sentiment analysis techniques are used. Sentiment analysis, or opinion mining, aims at user's attitude and opinions by investigating, analysing and extracting subjective texts involving users' opinions, preferences and sentiment. This is used particularly in data mining field for social media with many applications including product ratings and feedback analysis and customer decision making etc. Presence of emoticons, slang words and misspellings in tweets forced to have a pre-processing step before feature extraction.

There are different feature extraction methods for collecting relevant features from text which can be applied to tweets also. But the feature extraction is to be done in two phases to extract relevant features. In the first phase, twitter specific features are extracted. Then these features are removed from the tweets to create normal text. Again, feature extraction is done to get more features. This is the idea used in this paper to generate an efficient feature vector for analysing twitter sentiment. Since no standard dataset is available for twitter posts of electronic devices, we created a dataset by collecting tweets for a certain period. By doing sentiment analysis on a specific domain, it is possible to identify the influence of domain information in choosing a feature vector. Different classifiers are used to do the classification to find out their influence in this domain with this feature vector.

1.1. Problem Statement

Despite the availability of software to extract data regarding a person's sentiment on a specific product or service, organizations and other data workers still face issues regarding the data extraction. With the rapid growth of the World Wide Web, people are using social media such as Twitter which generates big volumes of opinion texts in the form of tweets which is available for the sentiment analysis. This translates to a huge volume of information from a human viewpoint which make it difficult to extract a sentence, read them, analyse tweet by tweet, summarize them and organize them into an understandable format in a timely manner.

Emoticons, are a pictorial representation of human facial expressions, which in the absence of body language and prosody serve to draw a receiver's attention to the tenor or temper of a sender's nominal verbal communication, improving and changing its interpretation. For example, ☺indicates a happy state of mind.

2. RELATED WORK

There are two basic methodologies to detect sentiments from text. They are Symbolic techniques and Machine Learning techniques. The next two sections deal with these techniques.

2.1 Symbolic Techniques

Turney used bag-of-words approach for sentiment analysis. In that approach, relationships between the individual words are not considered and a document is represented as a collection of words. To determine the overall sentiment, sentiments of every word is determined, and those values are combined with some aggregation functions. He found the polarity of a review based on the average semantic orientation of tuples extracted from the review where tuples are phrases having adjectives or adverbs. He found the semantic orientation of tuples using the search engine AltaVista.

Baroni et al. developed a system using word space model formalism that overcomes the difficulty in lexical substitution task. It represents the local context of a word along with its overall distribution.

Balahur et al. introduced EmotiNet, a conceptual representation of text that stores the structure and the semantics of real events for a specific domain. Emotinet used the concept of Finite State Automata to identify the emotional responses triggered by actions. In coarse grained approach, they performed binary classification of emotions and in fine grained approach they classified emotions into different levels.

2.2 Machine Learning Techniques

Machine Learning techniques use a training set and a test set for classification. Training set contains input feature vectors and their corresponding class labels. Using this training set, a classification model is developed which tries to classify the input feature vectors into corresponding class labels. Then a test set is used to validate the model by predicting the class labels of unseen feature vectors. Many machine learning techniques like Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) are used to classify reviews.

Some of the features that can be used for sentiment classification are Term Presence, Term Frequency, negation, n-grams and Part-of-Speech. These features can be used to find out the semantic orientation of words, phrases, sentences and that of documents. Semantic orientation is the polarity which may be either positive or negative.

Naive Bayes works well for certain problems with highly dependent features. Zhen Niu et al. introduced a new model in which efficient approaches are used for feature selection, weight computation and classification. The new model is based on Bayesian algorithm. Here weights of the classifier are adjusted by making use of 'Representative feature' which is the information that represents a class and 'Unique feature' is the information that helps in distinguishing classes. Using those weights, they calculated the probability of each classification and thus improved the Bayesian algorithm.

Barbosa et al. designed a 2-step automatic sentiment analysis method for classifying tweets. They used a noisy training set to reduce the labelling effort in developing classifiers. Firstly, they classified tweets into subjective and objective tweets. After that, subjective tweets are classified as positive and negative tweets. After doing normalization, they used probabilistic models to identify polarity lexicons. They performed classification using the BoosTexter classifier with these polarity lexicons as features and obtained a reduced error rate.

Wu et al. proposed a influence probability model for twitter sentiment analysis. If "@username" is found in the body of a tweet, it is influencing action and it contributes to influencing probability.

Pak et al. created a twitter corpus by automatically collecting tweets using Twitter API and automatically annotating those using emoticons. Using that corpus, they built a sentiment classifier based on the multinomial Naive Bayes classifier that uses "N-gram" and "POS-tags" as features. In that method, there is a chance of error since emotions of tweets in training set are labelled solely based on the polarity of emoticons. The training set is also less efficient since it contains only tweets having emoticons.

Xia et al. used an ensemble framework for sentiment classification. Ensemble framework is obtained by combining various feature sets and classification techniques. They used two types of feature sets and three base classifiers to form the ensemble framework. Two types of feature sets are created using POS information and Word-relations. Naive Bayes, Maximum Entropy and Support Vector Machines are selected as base classifiers. They applied different ensemble methods like Fixed combination, Weighted combination and Meta-classifier combination for sentiment classification and obtained better accuracy.

3. PROPOSED SOLUTION

Twitter posts of electronic products creates a dataset. Tweets are short messages with slang words and misspellings. So, we perform a sentence level sentiment analysis. This can be done in 3 phases. In first phase, pre-processing is done. Then a feature vector is created using relevant features.

Finally using different classifiers, tweets are classified into positive and negative classes. Based on the number of tweets in each class, the final sentiment is derived.

3.1. Creation of a Dataset

TABLE -1 Statistics of the dataset used

Dataset	Positive	Negative	Total
Training	500	500	1000
Test	100	100	200

Tweets are collected automatically using Twitter API and they are manually annotated as positive or negative. A dataset is created by taking 600 positive tweets and 600 negative tweets. Table-1 shows how dataset is split into training set and test set.

3.2. Pre-processing of Tweets

A pre-processing step is performed before feature extraction. Pre-processing steps include removing URL, avoiding misspellings and slang words. Misspellings are avoided by replacing repeated characters with 2 occurrences. Slang words contribute much to the emotion of a tweet. Hence, a slang word dictionary is maintained to replace slang words occurring in tweets with their associated meanings.

3.3. Creation of Feature Vector

Feature extraction is done in two steps. In the first step, twitter specific features are extracted. Hashtags and emoticons are the relevant twitter specific features. Emoticons can be positive or negative. So, they are given different weights. Positive emoticons are given a weight of "+1" and negative emotions are given a weight of "-1". There may be positive and negative hashtags. Therefore, the count of positive hashtags and negative hashtags are added as two separate features in the feature vector.

Twitter specific features may not be present in all tweets. So, a further feature extraction is to be done to obtain other features. After extracting twitter specific features, they are removed from the tweets. Tweets can be then considered as simple text. Then using unigram approach, tweets are represented as a collection of words. In unigrams, a tweet is represented by its keywords. A negative keyword list, positive keyword list and a list of different words are maintained that represent negation. Counts of positive and negative keywords in tweets are used as two different features in the feature vector. Presence of negation contribute much to the sentiment.

All keywords cannot be treated equally in the presence of multiple positive and negative keywords. Hence a special keyword is selected from all the tweets. In the case of tweets having only positive keywords or only negative keywords, a search is done to identify a keyword having relevant part of speech. Keywords that are adjective, adverb or verb shows more emotions. If a relevant part of speech can be determined for a keyword, then that is taken as special keyword. Otherwise a keyword is selected randomly from the available keywords as special keyword. If both positive and negative keywords are present in a tweet, we select any keyword having relevant part of speech. If relevant part of speech is present for both positive and negative keywords, none of them is chosen. Special keyword feature is given a weight of '1' if it is positive and '-1' if it is negative and '0' in its absence. Part of speech feature is given a value of '1' if it is relevant and '0' otherwise.

Thus, feature vector is composed of 8 relevant features. The 8 features used are part of speech (POS) tag, special keyword, presence of negation, emoticon, number of positive keywords, number of negative keywords, number of positive hash tags and number of negative hash tags.

3.4. Sentiment Classification

After creating a feature vector, classification is done using Naive Bayes, Support Vector Machine, Maximum Entropy and Ensemble classifiers and their performances are compared.

4. CLASSIFICATION TECHNIQUES

There are different types of classifiers that are generally used for text classification which can be also used for twitter sentiment classification.

4.1. Naive Bayes Classifier

Naive Bayes Classifier makes use of all the features in the feature vector and analyses them individually as they are equally independent of each other. The conditional probability for Naive Bayes can be defined

$$P(X|y_j) = \Pi_{i=1}^{m} P(x_i|y_j) \qquad (1)$$

as 'X' is the feature vector defined as:

X=fx$_1$, x$_2$....x$_m$ g and y$_j$

Here, in our work there are different independent features like emoticons, emotional keyword, count of positive and negative keywords, and count of positive and negative hash tags which are effectively utilized by Naive Bayes classifier for classification. Nave Bayes does not con-sider the relationships between features. So, it cannot utilize the relationships between part of speech tag, emotional keyword and negation.

4.2. SVM Classifier

SVM Classifier uses large margin for classification. It separates the tweets using a hyper plane. SVM uses a

$$g(X) = w^T \phi(X) + b \qquad (2)$$

discriminative function defined as

’X’ is the feature vector, ’w’ is the weights vector and ’b’ is the bias vector. () is the nonlinear mapping from input space to high dimensional feature space. ’w’ and ’b’ are learned automatically on the training set. Here a linear kernel is used for classification. It maintains a wide gap between two classes.

4.3. Maximum Entropy Classifier

$$P_\lambda(y|X) = 1/Z(X)exp\left\{\sum_i \lambda_i f_i(X,y)\right\} \qquad (3)$$

In Maximum Entropy Classifier, no assumptions are taken regarding the relationship between features.

$$f_i(X,y) = \begin{cases} 1, & \text{X=x}_i \text{ and } y = y_i \\ 0, & \text{otherwise} \end{cases} \qquad (4)$$

This classifier always tries to maximize the entropy of the system by estimating the conditional distribution of the class label. The conditional distribution is defined as ’X’ is the feature vector and ’y’ is the class label. Z(X) is the normalization factor and $_i$ is the weight coefficient. f_i (X,y) is the feature function. In feature vector, the relationships between part of speech tag, emotional keyword and negation are utilized effectively for classification.

4.4. Ensemble Classifier

Ensemble classifiers can be of different types. They try to make use of the features of all the base classifiers to do the best classification. The base classifiers used here are Nave Bayes, Maximum entropy and SVM. Here an ensemble classifier is generated by voting rule. The classifier will classify based on the output of majority of classifiers.

5. EVALUATION

Since we have selected product domain, there is no need of analysing subjective and objective tweets separately. To identify the quality of product, both qualities contribute similarly. This shows how context or domain information affects sentiment analysis. These classifiers are tested using MATLAB simulator.

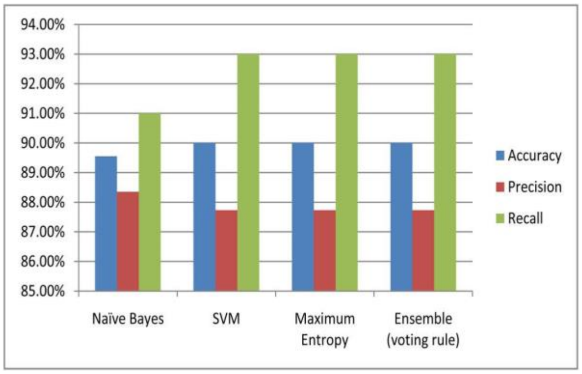

Fig -1. Performance of Different classifiers in Twitter Sentiment Analysis

We used three types of basic classifiers (SVM, Nave Bayes, Maximum Entropy) and ensemble classifier for sentiment classification. SVM and Naive Bayes classifiers are implemented using MATLAB built in functions. Maximum Entropy classifier is implemented using MaxEnt software. All these classifiers have almost similar performance.

Naive Bayes has better precision compared to the other three classifiers, but slightly lower accuracy and recall. SVM, Maximum Entropy Classifier and Ensemble classifiers have similar accuracy, precision and recall. They obtained an accuracy of 90% whereas Naïve Bayes has 89.5%. This shows the quality of the feature vector selected for the product domain. This feature vector aids in better sentiment analysis despite of the classifier selected.

6. CHALLENGES

6.1. Incremental Approach

Whenever data is added we need to do analysis for which we can use the previous analysis result. Incremental approach allows an existing result to be updated using only new individual data instances, without having to re-process past instances which is useful in situations where the entire dataset is not available when the data changes over time.

6.2. Credibility/Behaviour/Homophily

Behaviours in social media are only observed by the traces they leave in social media. Even if a behaviour is analysed on social media and related patterns are gleaned, it's difficult to verify the validity of these behavioural patterns. Evaluation becomes even more challenging for industries in which important decisions are to be made based on observations of individual behaviour.

6.3. Sarcasm

Sarcasm can be used to hurt or offend or can be used for comic affect. It is an ironic or satirical remark that seems to be praising someone or something but is really taunting or cutting. It means false positives for eg. "Children really brighten up a household - they never turn the lights off". Detecting sarcasm from the expressions and finding out the correct context related sentiments is a challenging task.

6.4. Parallel Computing for Massive Data

If we divide the computation into tasks or processes that can be executed simultaneously, then there can be an improvement in the speed using parallelism, it is necessary to achieve in sentiment analysis for massive data of social media, where massive instant messages are published every day so that we can utilize the overall computing power.

6.5. Review Author Segmentation

Opinion towards a target may be specified by many people who can be called as review authors. Depending on the commenting style of these authors, they should be categorized so that credibility evaluation will be easy. In decision making this credibility evaluation is helpful.

6.6. Refinement of Existing Lexicons or Updating/ Downdating Lexicons

Many people comments, the Performance of sentiment analyser depend on the correctness of the lexicon. Finetuning of existing lexicons is required to accommodate new words and destroy the words which are no more used for better results. Lexicon expansion using synonyms has a drawback of the wording losing it primary meaning after a few recapitulations.

6.7. Grammatically Incorrect Words

There are many approaches that analyse sentiments but hardly any work accomplished on grammatical errors. The results of sentiment analysis can be improved if these types of errors can be mapped to correct words.

6.8. Handling Noise and Dynamism

Social media data are enormous, noisy, unstructured, and dynamic in nature, and thus novel challenges arise, introduces representative research problems of mining social media. Identifying and removal of noisy data is a challenging task.

7. SYSTEM ARCHITECTURE

The methodology over which our system is based on is as follows:

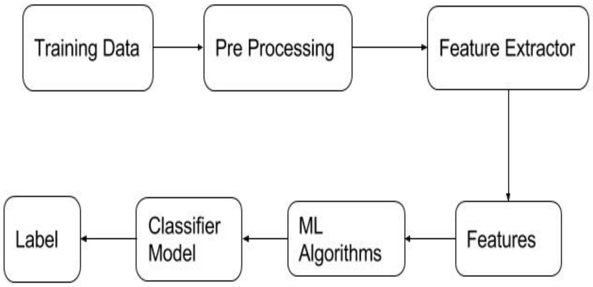

Fig -2: Methodology followed by the system

We use different feature sets and machine learning classifiers to determine the best combination for sentiment analysis of twitter. We also experiment with various pre-processing steps like - punctuations, emoticons, twitter specific terms and stemming. We investigated the following features - unigrams, bigrams, trigrams and negation detection. We finally train our classifier using various machine-learning algorithms - Naive Bayes, Decision Trees and Maximum Entropy. We present a new feature vector for classifying the tweets as positive, negative and extract peoples' opinion about products.

So, the architecture which our system is following can be represented by:

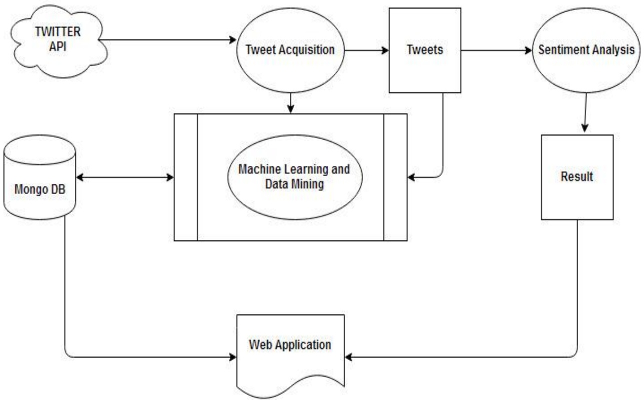

Fig -3: System Architecture.

8. RESULT AND DISCUSSION

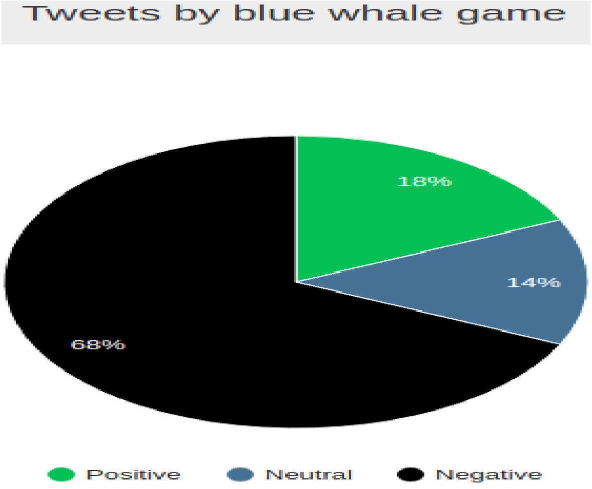

Fig -4: Result obtained based on the analysis done.

8.1 Twitter Retrieved

Developer needs to agree in terms and conditions of development Twitter platform which has been provided to get an authorization to access a data. The output will be saved in JSON file. JSON (JavaScript Object Notation) is a lightweight data-interchange format which is easy for humans to write and read. JSON is simple for machines to generate and parse. JSON is a text format that is totally language independent. However, output's size depends on the time for retrieving tweets from Twitter. The output will be categorized into encoded and unencoded. The tweets will be assigned the value of each word, together with categorize into positive and negative word. The result will be shown in .txt, .csv and html.

8.2. Information Presented

The result will be shown in a pie chart which is representing a percentage of positive, negative and null sentiment hash tags. Null hash tag represents the hash tags that were assigned zero value. However, this program can list a top ten positive and negative hash tags.

8.3 Sentiment Analysis

Tweets from JSON file will be assigned the value of each word by matching with the lexicon dictionary. As a limitation of words in the lexicon dictionary which is not able to assign a value to every single word from tweets. However, as a scientific language of python, which can analyse a sense of each tweet into positive or negative for getting a result.

9. CONCLUSIONS

There are different Symbolic and Machine Learning techniques to identify sentiments from text. Machine Learning techniques are simpler and efficient than Symbolic techniques. These techniques can be applied for twitter sentiment analysis. An efficient feature vector is created by doing feature extraction in two steps after proper pre-processing. In the first step, twitter specific features are extracted and added to the feature vector. After that, these features are removed from tweets and again feature extraction is done as if it is done on normal text. These features are also added to the feature vector. Classification accuracy of the feature vector is tested using different classifiers like Nave Bayes, SVM, Maximum Entropy and Ensemble classifiers. All these classifiers have almost similar accuracy for the new feature vector. This feature vector performs well for electronic products domain.

ACKNOWLEDGEMENT

The project "Sentiment Analysis in Twitter" was successfully implemented with the help of our Project Guide Prof. Dr, D. R. Ingle and Project Coordinator Prof. Rahul Patil. We would like to thank them for their valuable guidance along with their continuous encouragement and support whenever required. We express our gratitude to Principal Dr. M. Z. Shaikh for his coordination and obliging us with his great knowledge.

REFERANCES

[1] Fuji Ren, Ye Wu," Predicting User-Topic Opinions in Twitter with Social and Topical Context", IEEE Trans. on affective computing, vol. 4, no. 4, October-December 2013

[2] Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li, "Dual Sentiment Analysis: Considering Two Sides of One Review", IEEE Trans.on Knowledge and Data Engineering, 2015

[3] Danushka Bollegala, David Weir, and John Carroll," Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus", IEEE trans. on knowledge and data engineering, vol. 25, no. 8, August,2013.

[4] A. K. Jose, N. Bhatia, and S. Krishna "Twitter Sentiment Analysis". National Institute of TechnologyCalicut,2012

[5] Y. Wu and F. Ren, "Learning sentimental influence in twitter," in Future Computer Sciences and Application

(ICFCSA), 2011 International Conference on, pp. 119–122, IEEE, 2011.

[6] J. Spencer and G. Uchyigit, “Sentiment or: Sentiment Analysis of Twitter Data,” Second Joint Conference on Lexicon and Computational Semantics. Brighton: University of Brighton, 2013.

[7] B. Pang, and L. Lee, “Opinion mining and sentiment analysis,” 2nd workshop on making sense of Microposts. Ithaca: Cornell University. Vol.2(1),

[8] G. Vinodhini and R. Chandrasekaran, “Sentiment analysis and opinion mining: A survey,” International Journal, vol. 2, no. 6, 2012.