International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 08 | Aug -2017 www.irjet.net p-ISSN: 2395-0072

Hybrid Classifier for Sentiment Analysis Using Effective Pipelining

Akhil Sharma[1], Aman Sharma[2], Rajeev Kumar Singh[3], Dr. Madhur Deo Upadhayay[4]

[1,2]*Electronics and Communication Engineering, Shiv Nadar University, Uttar Pradesh, India*
[3]*Assistant Professor, Dept. of Electrical Engineering, Shiv Nadar University, Uttar Pradesh, India*
[4]*Assistant Professor, Dept. of Computer Science, Shiv Nadar University, Uttar Pradesh, India*

Abstract - *A Social media has become a platform for people to express their thoughts, opinions and ideas. Facebook, Twitter, Google+ and the likes have emerged as data hubs for people wanting to improve market sales, predict outcomes of events, and characteristics of human behavior. Polling and surveys are outdated and lengthy techniques. With opinion mining and sentiment analysis data extraction and classification becomes easy. In this paper, we have used a hybrid method for analyzing sentiments. This method employs a pipeline system consisting of rules, lexicon and machine learning based classifier where a tweet after undergoing preprocessing is first classified by the lexicon and the rules classifier and is sent to the machine learning module only if the tweet's analysis score doesn't achieve a predetermined threshold value. A comparison is made between the individual - rules, lexicon, and machine learning approaches, and hybrid classifier on the basis of F-score, recall and precision.*

Key Words: **opinion mining, sentiment analysis, rules-based, lexicon-based, classifier, hybrid approach**.

1. INTRODUCTION

Twitter is one of the most popular microblogging and social networking websites [4]. People from time to time post on Twitter, an activity called tweeting. The diversity of people on twitter makes the tweets more versatile and valuable [7]. Therefore, Twitter becomes one of the most valuable places to find opinions on any issue. This allows computer scientists to perform credible sentiment analysis and develop pathways for data mining. This data can be used in marketing, sales or poll analysis. Timely feedback on products can be collected by evaluating people's tweets on Twitter [1,2,3].

Researchers can use the data sets to build unsolicited public opinion polls on important social matters [1]. Social media becomes a powerful tool for common public to get involved with politics, media and business intrinsically. Polls are expensive and time consuming [1,2]. With continued improvement in data analysis techniques, these tasks have become practically viable. The credibility of data and results is higher than before. Manual surveys and polls are not always trustable, whereas there is significantly less or negligible scope for human errors in data mining and subsequent analysis. Political inclinations, interests of common public will be available for parties to understand and prepare for their campaigns. The needs of people and complains from the society will become accessible to politicians. The gap between the government and public can be bridged with ease. Predictions pertaining to elections or major events can also be extracted in one go [1].

After any incident, protest or social unrest, people log into social media websites to post or to make a comment in order to express their thoughts and opinions. Social media is powerful in terms of spreading social awareness about crimes, diseases, and other epidemics. Twitter has become a solid and trustable commodity not only for its users but also researchers. The data consolidated can give great pictorial trends regarding people's opinions. The unprecedented view of public is displayed on social media, especially on Twitter [1].

Sentiment analysis is a field of study to find how sentiments and opinions are expressed in texts. Approaches that are used to classify sentiments include - rules based, lexicon based, machine learning and using deep learning techniques [2,3,10,11]. The method of classifying tweets on the basis of pre-fixed rules is called rules based approach. The approach of using opinion words or the lexicon to determine opinion orientations is called lexicon based approach [1,5]. Rules based approach along with lexicon based approach has high precision but low recall [2]. Emoticons, informal language and abbreviations are some of the parts of unstructured textual data that may go undetected or unclassified in the lexicon based approach. For example "Mauritius is a gr8 holiday destination," is a sentence of positive demeanor. However, a classifier using lexicon based approach might classify it as neutral or no. Although, it is possible to add these expressions in the opinion lexicon, due to continuous change in their usage, it becomes hard to classify [2].

Another method that is used for sentiment analysis is the machine learning approach [4,8]. This method is effective for classification of sentences and documents by training the classifier to determine positive, negative and neutral sentiments [4,8]. Since manual labelling of large set of tweets is often time consuming and difficult, this approach is not easy to implement [2]. Also, Deep Learning algorithms could provide the most accurate results, but these techniques are extremely computationally expensive to train. To optimize the large amount of matrix multiplication operations that deep learning involves, substantial investment is needed to upgrade the IT infrastructure for more processing power. Moreover, deep learning requires immense amount of data to train the model as compared to traditional machine

learning algorithms. The execution time to train a deep learning algorithm is significantly more than usual, taking days and weeks on end. Since we are trying to devise the best solution that optimizes processing speed, accuracy and execution time, deep learning is not the best standalone solution to implement.

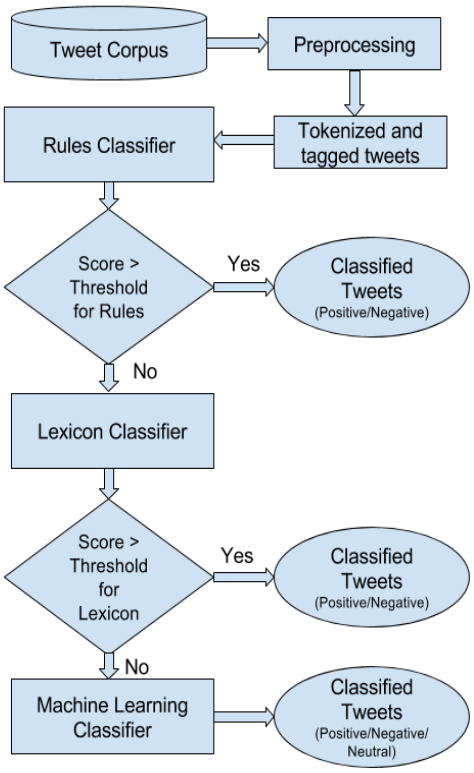

Figure 1: Architectural overview of the algorithm

In this paper, we discuss a hybrid approach in which the individual modules of rules, lexicon and machine learning classifiers have been pipelined in a way that results in better optimization in terms of performance and speed. The hybrid system is divided into stages with the first two stages being rules classifier and lexicon classifier. The output of the initial preprocessing system is passed through the hybrid model. Rules classifier, being the first stage of the model, tries to classify the tweets. The tweets that exceeds a certain confidence threshold, exit from the hybrid model from this step itself. The rest encounter the next step, i.e., lexicon based method. Tweets passing this threshold test don't advance to the machine learning stage and the classification of the lexicon classifier is the accepted output. The tweets that weren't able to achieve the set threshold were then finally passed to the machine learning classifier which uses the SVM (Support Vector Machine) algorithm to classify a tweet.

2. RELATED WORKS

This paper covers the study of sentiment analysis and opinion mining. We use various approaches to determine whether a sentence, statement or document is positive, negative or neutral. As discussed, the three main approaches are: lexicon based, machine learning based and the hybrid approach.

The lexicon-based approach (Ding et al., 2008, Taboada, et al., 2010) determines the polarity and sentiment of any given statement using opinionated words. (Ding et al., 2008) proposed a new approach which instead of looking at the current sentence alone, exploited external information and evidences in other sentences and reviews, and some linguistic orientations of opinion words [5]. (Saif et al., 2015) proposed a method to take contexts to capture their semantics and update their pre-assigned strength and polarity in the lexicon [12]. As mentioned earlier, this method results in high precision but low recall.

The machine learning based approach trains a sentiment classifier using features such as N-grams (Pak et al., 2010). In their paper, they compared different learning techniques such as SVM (Support Vector Machines), Maximum Entropy, Naïve Bayes, etc. (Pak et al.,2010) used a classifier based on the multinomial Naïve Bayes classifier that uses N-gram and POS-tags as features [4]. The third approach is the hybrid approach which combines both the lexicon based and machine learning approach. This idea was also applied to sentiment classification of reviews in (Tan et al., 2008) which classified reviews into two classes, positive and negative, but no neutral class making the problem much easier. (Zhang et al., 2011) proposed a classification system working with ternary sentiment values; positive, negative and neutral. A similar approach was used by (Pedro P. Balage Filho et al., 2014) in SemEval-2014.

3. TEXT PREPROCESSING

Twitter textual data in its raw form is an unstructured form of data upon which data analysis techniques could not be applied directly. After collection of tweets, the data goes through various steps in order to clean tweets after addressing several challenges that the twitter data poses. Throughout the process, we tokenized the tweets, filtered the unrequired terms, removed the stop words and afterwards applied stemming and lemmatization methods to the tokens.

3.1 Tokenization

Tokenization is basically the process of splitting a stream of text into smaller entities, usually words or phrases, as per the predefined rules. This is an important step in text analysis, although a basic one. Initially, we tried to implement this using the nltk library's tokenize function. However this general-purpose English tokenizer does not capture peculiarities such as emoticons, @, URLs, #hashtags [13]. Therefore, we used regular expressions that addressed

the specific case of Twitter data. The user defined regular expressions identified the atypical entities and ignored them in individual tokens. [13] They were separated out from the main data which solved this problem. For example a tweet "@POTUS what's the government doing to solve the Syrian crisis" when tokenized using the nltk library creates two tokens for "@" and "POTUS" while using regular expressions only a single token is created, i.e., "@POTUS".

3.2 Stop word removal

Some words in every language are common in sentences, but hold little meaning when it comes to usage. Without the contextual peripherals, these words mean nothing [13]. In English language such words are articles, conjunctions, adverbs etc. These are called stop-words. Stop-word removal is an important step during the pre-processing stages. We used the list provided by the nltk library although a custom list of stop words could also be built [13]. We also removed punctuation marks and terms like *RT* (used for re-tweets) and *via* (used to mention the original author of an article or a retweet), which are not included in the default stop-word list [2,13].

3.3 Stemming and lemmatization

For purpose of grammatical correctness and contextual cohesiveness, documents use different forms of a word, such as *organize*, *organizes*, and *organizing*. Additionally, there are families of derivationally related words with similar meanings, such as *democracy*, *democratic*, and *democratization* [5]. The goal of stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form [5]. For instance:

am, are, is $\Rightarrow$ be

car, cars, car's, cars' $\Rightarrow$ car

The result of this mapping of text will be something like:

the boy's cars are different colors $\Rightarrow$ the boy car be differ color [5]

We used WordNetLemmatizer and PorterStemmer functions of nltk library for lemmatizing and stemming purposes respectively.

4. SYSTEM ARCHITECTURE

4.1 Normalization and Rule-based Classifier

The Rules based classifier module primarily looks for capturing the predominant sentiment of a tweet by focusing on emoticons present in the tweet text. Since people generally tend to use emoticons to express their overall mood, they could be treated as a sole representor of the tweet, irrespective of the tweet text. Since the data received from preprocessing step are plain tokens, normalization is required that could provide some contextual information about the tokens and aid in efficient classification. For the same reason, hashtags, user mentions and URLs are changed to text format.

For normalizing and tagging the texts we use a normalization module that performs the following operations:

- Hashtags, URLs and user mentions are converted into codes [3]. These codes are basically the textual representation of their respective symbols.
- Emoticons are grouped into categories like 'happy','sad', 'laugh' and are converted to particular codes that best represent the emotions in broad categories [3]
- Part-of-speech tagging using Ark-twitter NLP (Brendan O'Connor et al., 2013) to give POS tags to each token.

4.2 Lexicon based Classifier

The lexicon-based classifier is based on the idea that the polarity of a text can be given by the sum of the individual polarity values of each word or phrase present in the text [3]. We used the sentiment lexicon provided by Opinion-Lexicon (Hu and Liu, 2004) and a list of sentiment hashtags provided by the NRC Hashtag Sentiment Lexicon (Mohammad et al., 2013). To deal with negation occurring in sentences, we built a list of negative words. Negating words completely change the sentence's sentiment upon combining with any token with polar sentiment value. Our classifier assigns polarities to each word which are then added up to give the overall polarity of the text [3]. We searched for individual tokens in the lexicon and only the words that are found are assigned the polarity. The classifier sets value as +1 if they are positive words and -1 if they are negative. When a negating word is found in the text, the overall value of the word is inverted.

4.3 Machine Learning Classifier

In the field of sentiment analysis, mainly the supervised methods of machine learning are used. Among the supervised algorithms, we used SVM (Support Vector Machine) as it outperforms other machine learning classifiers and gives better results in high dimensional feature space. We provide labeled data (training data) to the machine learning classifier. This training data acts as a fuel for the classifier from which it can predict the sentiment of the new data it will receive (testing data). We use various features such as N-grams, presence of negation, frequencies of positive/negative words etc. that are extracted from the input data.

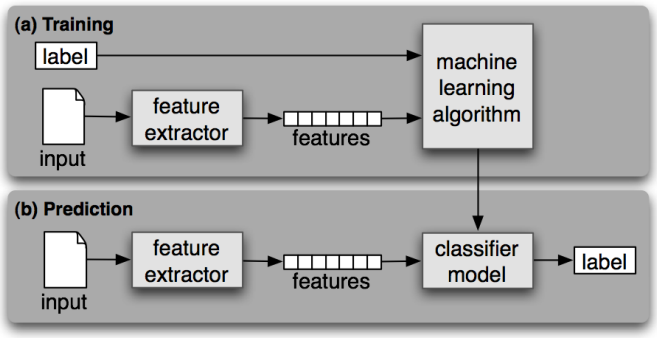

Figure 2: Machine learning approach [15]

4.4 Hybrid Approach

The hybrid approach combines the rules based, lexicon based and machine learning approach into a unified system that utilizes the strong areas of the individual classifiers while simultaneously trying to avoid their shortcomings. Rules based and Lexicon classifiers have high precision but low recall. Hence, they act as the initial two stages of the Hybrid System. If the tweets contain emoticons and we assume that no sarcasm is intended, rules based approach performs fairly well in capturing the tweets with explicit emoticons early in the analysis stage. This results in shorter classification time by bypassing further stages. Lexicon classifier with its high precision comes next in the line. The unclassified tweets from the rules based classifier encounter Lexicon classifier. If the message contains opinionated words, it can effectively be classified by this stage itself. SVM based machine learning classifier is the third and final stage that classifies the tweet as positive, negative or neutral based on the training dataset. The accuracy of the classifier is directly correlated to the selection of appropriate features. By pipelining the system in this way, we were able to achieve good accuracy and shorter processing time with machine learning stage coming into picture only for the messages left unclassified by earlier two stages.

5. RESULTS

5.1 UMICH S1650 – Sentiment Classification

We classified two datasets as discussed earlier. Now we will evaluate the performance of our classifier on the basis of F-score, Recall, Precision and Accuracy on the UMICH SI650 - Sentiment Classification dataset.

5.1.1 Hybrid Classifier

Table 1 – Hybrid classifier's performance for UMICH SI650

Sentiment	Recall	Precision	F-score
Positive	78.37	92.03	83.21
Negative	90.77	76.82	84.652

The hybrid classifier achieved an accuracy of 83.76%. Since the test set only contained positive and negative sentiments, there is no score for neutral sentiments. For the entire testing dataset, our algorithm classified 5 examples (0.07% of the dataset) using the rule-based classifier, and 3873 examples (54.65% of the dataset) using the lexicon-based classifier. Since the machine learning classifier had no thresholding, it classified every message. Only the messages that were not classified by either the lexicon classifier or the rule classifier (3208, 45.27% of the dataset) encountered the machine learning classifier.

5.1.2 Rule-based Classifier

Table 2 – Rules-based classifier performance for UMICH SI650

Sentiment	Recall	Precision	F-score
Positive	0.20	91.66	0.39
Negative	0.17	87.31	0.34

The results in Table 2 are for the data that was classified using the rule-based classifier only. In order to increase the accuracy of the hybrid classifier, we have introduced threshold values. The rule-based classifier would be able to classify the message only when the score of message goes above these threshold values. In case if it fails, the lexicon classifier will be called. The values for threshold have been set empirically for the two stages to get the best possible results. This improves the individual classifier modules as well as the performance of the hybrid system altogether.

5.1.3 Lexicon Classifier

Table 3 – Lexicon classifier's performance for UMICH SI650

Sentiment	Recall	Precision	F-score
Positive	70.76	94.01	80.74
Negative	81.04	86.73	83.78

The results in Table 3 above are for the data that was classified using the lexicon classifier. The lexicon classifier is able to classify a large subset of the tweets with good recall and a high precision. Similar to the case of rules classifier, in order to increase the accuracy of the hybrid classifier we have introduced threshold values in the lexicon stage. If the score exceeds these set threshold values, the lexicon classifier would be able to classify the message and in case if it fails, the machine learning classifier will be called. Once the threshold values are applied, the classifier's accuracy becomes 94.52% (increased by 15.04%). A significant improvement in recall, precision and F-score of the classifier is also noticed.

5.1.4 Machine Learning Classifier

Table 4 – Results for Machine learning classifier for UMICH SI650 dataset

Sentiment	Recall	Precision	F-score
Positive	85.98	89.52	87.71
Negative	52.89	93.16	67.47

Inspecting the results from Table 4, we see that the machine learning classifier performs better than lexicon classifier for positive sentiment but lags behind in detecting negative ones. Since this was the last stage of the hybrid classifier, no threshold values have been kept for the scores obtained by messages. Table 5 summarizes the results obtained by each individual classifier and by the hybrid classifier in classifying messages from the test dataset. In the task, the systems were evaluated with the average F-score obtained for positive and negative classes. The hybrid approach performs better than the rest three individual classifiers by slightly outperforming the lexicon based approach.

Table 5 – Performance summary of various classifiers based on average F-score

Classifier	Twitter2014 Test dataset (F-score)
Rule-based	0.26
Lexicon-based	82.26
0Machine learning approach	77.59
Hybrid approach	83.91

5.2 SemEval-2014 Task 9: Sentiment Analysis in Twitter

SemEval (Semantic Evaluation) is a series of evaluations of computational semantic analysis systems, organized under the umbrella of SIGLEX, the Special Interest Group on the Lexicon of the Association for Computational Linguistics. The second dataset that we used for classification was from SemEval-2014 Task 9: Sentiment Analysis in Twitter which required classification of messages into ternary classes - positive, negative and neutral.

5.2.1 Hybrid Classifier

The hybrid classifier achieved an accuracy of 55.78%. For the entire testing dataset, our algorithm classified 344 examples (3.38% of the dataset) using the rule-based classifier, and 5183 examples (50.96% of the dataset) using the lexicon-based classifier. The machine learning classifier without any thresholding, was the last stage in the hybrid system. It classified the messages that could not be classified by either the lexicon classifier or the rule-based classifier (4643, 45.65% of the dataset). In the next few subsections we will discuss the performance of individual classifiers- Rules based, Lexicon based and Machine learning classifier.

Table 6 – Hybrid classifier's performance in SemEval-2014's dataset

Sentiment	Recall	Precision	F-score
Positive	62.14	79.71	69.83
Negative	44.17	68.78	53.79
Neutral	61.69	9.4	16.31

5.2.2 Rule-based Classifier

Table 7 – Rules based classifier's performance in SemEval-2014's dataset

Sentiment	Recall	Precision	F-score
Positive	9.01	83.40	16.26
Negative	1.98	70.75	3.85
Neutral	96.46	5.77	10.89

5.2.3 Lexicon Classifier

Table 8 – Lexicon based classifier's performance in SemEval-2014's dataset

Sentiment	Recall	Precision	F-score
Positive	58.98	81.79	69.83
Negative	41.05	70.36	51.85
Neutral	64.20	9.15	16.02

5.2.4 Machine Learning Classifier

Table 9 – Performance of Machine learning classifier in SemEval-2014's dataset

Sentiment	Recall	Precision	F-score
Positive	63.64	79.44	70.66
Negative	38.55	79.15	51.84
Neutral	68.88	9.25	23.69

The table below shows the results obtained by each individual classifier as well as the hybrid classifier in classifying messages in the test dataset. In the task, the systems were evaluated with the average F-score obtained for positive and negative classes.

Table 10 – Performance summary based on average F-scores of each classifier

Classifier	Twitter2014 Test dataset (F-scores)
Rule-based	10.05
Lexicon-based	60.84
Machine learning approach	61.25
Hybrid approach	61.81

One of the top SemEval leaderboard scores that didn't involve deep learning algorithms was 60.83, achieved by (P.B. Filho et al., 2014) using a hybrid approach with machine learning classifier. Whereas, we were able to achieve a score of 61.81 through our hybrid classifier employing efficient pipelining.

6. CONCLUSION

The hybrid classifier proposed in this paper not only improves the accuracy but also achieves significant breakthrough in decreasing the GPU processing power. The processing time for end to end tweet classification also registered, on average, a decrease of 35.19% as compared to the machine learning classifier. This was achievable since there was nearly 45-50% reduced requirement for matrix computations involved in machine learning approach. Utilizing lexicon and rules based classifiers early in the hybrid system took some burden off the machine learning stage, thus delivering substantial improvements in performance and time. Also, since this is a modular system, refining the individual modules could further improve accuracy.

REFERENCES

[1] E. M. Cody, A. J. Reagan, P. S. Dodds, and C. M. Danforth, "Public Opinion Polling with Twitter," The University of Vermont, Aug. 2016.

[2] L. Zhang, R. Ghosh, M. Dekhil, M., and B. Liu, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis," HP Laboratories, 2011.

[3] P. B. Filho, L. Avanço, T. Pardo, and M.D.G.V. Nunes, "NILC_USP: An Improved Hybrid System for Sentiment Analysis in Twitter Messages," Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014.

[4] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," 2010.

[5] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," Proceedings of the international conference on Web search and web data mining - WSDM '08, 2008.

[6] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 Task 4: Sentiment Analysis in Twitter," Proceedings of the 10th Interntional Workshop on Semantic Evaluation (SemEval-2016), 2016.

[7] E. Palogiannidi, A. Kolovou, F. Christopoulou, F. Kokkinos, E. Iosif, N. Malandrakis, H. Papageorgiou, S. Narayanan, and A. Potamianos, "Tweester at SemEval-2016 Task 4: Sentiment Analysis in Twitter Using Semantic-Affective Model Adaptation," Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016.

[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?," Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02, 2002.

[9] F. Sommar and M. Wielondek, "Combining Lexicon- and Learning-based Approaches for Improved Performance and Convenience in Sentiment Classification," KTH ROYAL INSTITUTE OF TECHNOLOGY, 2015.

[10] A. Mudinas, D. Zhang, and M. Levene, "Combining lexicon and learning based approaches for concept-level sentiment analysis," Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '12, 2012.

[11]B. Lu and B. K. Tsou, "Combining a large sentiment lexicon and machine learning for subjectivity classification," International Conference on Machine Learning and Cybernetics, 2010.

[12] H. Saif, Y. He, M. Fernandez, and H. Alani, "Contextual semantics for sentiment analysis of Twitter," Information Processing & Management, vol. 52, no. 1, pp. 5–19, 2016.

[13]M Bonzanini, "Mastering Social Media Mining with Python: Acquire and Analyze Data from All Corners of the Social Web with Python". Birmingham: Packt, 2016.

[14]"Stemming and lemmatization," Stemming and lemmatization. [Online]. Available: http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html. [Accessed: 08-Mar-2017].

[15]" Learning to Classify Text," Learning to Classify Text. [Online]. Available: http://www.nltk.org/book/ch06.html. [Accessed: 08-Mar-2017].