Sentiment Analysis of Twitter Data Using Sentiment Influencers

Munazza Ishtiaq
College of Electrical & Mechanical Engineering
National University of Sciences & Technology (NUST)
Islamabad, Pakistan
munazza.ishtiaq@gmail.com

Abstract: *Sentiment analysis has attained much attention in the recent years due to its significance in various fields as it captures and analyzes such attitudes and opinions in an automated and structured fashion and offers a powerful technology to a number of problem domains. This research is based on the use of a novel unsupervised approach for sentiment analysis of twitter data using a rule based scoring engine. The focus of this approach is on POS tagging in which the parts-of-speech are ranked according to their sentiment describing influence. A novel term is devised for POS which is named as "sentiment influencers" and are ranked according to their influence on detecting sentiment. Results have shown that appropriate ranking of POS provides good results than its normal usage.*

Keywords: Component, Parts-of-speech (Pos), Rule Based Scoring Engine, Sentiment Analysis, Sentiment Influencers, Twitter

Recieved: 20 November 2014, Revised 18 December 2014, Accepted 23 December 2014

1. Introduction

Sentiment analysis is not a new field; a lot of research has been carried out in this field. The issue due to which sentiment analysis came into existence is what do people feel or think about a particular topic. There are many documents that present author's subjective views on particular topics. Sentiment analysis is the extraction of attitudes and opinions from human authored documents. Most of the researchers defined sentiment analysis as positive or negative opinion [1], [2], [3]. Sentiment analysis has attained much attention in the recent years due to its significance in various fields as it captures and analyzes such attitudes and opinions in an automated and structured fashion and offers a powerful technology to a number of problem domains, including business intelligence, marketing, national security, crime prevention and healthcare/wellbeing services. Social media in particular becomes an interesting and practical source for sharing sentiments, emotions and opinions therefore they are also useful for sentiment mining. The text written by any author is spontaneous, unstructured and disordered. There is a need to classify this text and to acquire some significant outcomes such as whether the author is admiring or criticizing. Sentiment analysis can also be used to filter emails and other messages, or indicate abusive messages in newsgroup or helps users navigate via the Internet not only using topic keywords but also opinions.

Social media plays an important role in any aspect of life. Social media can be defined as the social instrument for communication [4]. According to a survey held in 2014, 74% of online adults use social networking sites [5]. Social media allows users to interact

over the web and share their contents, ideas, photos, videos, send emails etc. Although social media is growing rapidly, still there is a need to get useful information from the data scattered on social media. All the data available on social media or social networking sites is raw, unstructured and spread all over. This research is based on sentiment analysis of twitter data. Twitter, a microblogging site has emerged as a popular social media site that allows its users to send and read the posts of up to 140 characters known as "Tweets". The main issues in extracting sentiments from tweets are: abbreviations, lack of capitals, poor grammar, poor punctuations, and poor spellings. Twitter is selected for this research due to its limited amount of characters in which users can precisely describe their thoughts thereby handling the associated issues.

This research is based on the extraction of sentiments in the form of positive tweets and negative tweets using a rule based scoring engine in which the focus is on the ranking of sentiment influencers. For this purpose, noun, verb, adverb and adjectives are considered as the main sentiment influencers. Nouns are mostly neutral because they are used to describe some person, thing, place or idea. Verbs describe the actions of the nouns and play a significant role in determining the polarity of a text for sentiment analysis. There is also a possibility that a sentence may only contain verb if no other POS is present in the sentence. Adverbs modify verbs and adjectives and are used as intensifiers. Adjectives modify nouns. Therefore, by understanding the concepts of these sentiment influencers, they are ranked and assigned scores. The proposed framework consists of five main phases: Data Acquisition, Pre-processing, Sentiment Influencers Identification, Rule based scoring engine focusing on sentiment influencers ranking and the final phase is the sentiment classifier which classifies the tweet as positive, negative or neutral. Details of the proposed framework will be discussed in the later section. Rest of the paper is arranged as under:

In section 2, related research in sentiment analysis has been discussed in general. In section 3, complete proposed framework has been discussed with details. Section 4 will cover details about implementation of the framework followed by validation of experimental results on twitter dataset in section 5. Finally Section 6 will conclude the paper and directions for future work will be discussed.

2. Related Work

Turney [6] developed a simple unsupervised learning technique for classification of reviews. He used the terms recommended (thumbs up) or not recommended (thumbs down) in his research. Pang [3] determined the sentiments of a document by overall text sentiments not only by topics. He used movie reviews for this purpose. He used three machine learning techniques that are Naive Bayes classification, maximum entropy classification, and support vector machines to test the results. Benamara et al. [7] have proposed AAC (Adverb Adjective Combination) sentiment analysis technique. They defined a general set of axioms and proposed three different AAC scoring methods that are Variable scoring, Adjective priority scoring (APS), and Adverb First Scoring (AdvFS) algorithms. The focus of their research was adverbs of degree. Kevin Gimpel et al. have developed a part-of-speech tagger for Twitter [8]. They developed a tag set, annotated data, developed features and reported results. They developed annotations in three different stages, introduced a simple and easy to apply POS inventory for twitter. They compared their developed system with the Stanford tagger. Hatzivassiloglou and McKeown[9] hypothesize that adjectives separated by \and" have the same polarity, while those separated by \but" have opposite polarity. Agarwal et al. carried out sentiment analysis of twitter data in [10]. They introduced a POS specific prior polarity feature and used a tree kernel to avoid the need for monotonous feature engineering. They created models for two classification tasks: firstly, a binary task for classifying tweets into positive or negative, secondly, a 3 way task to classify tweets as positive, negative or neutral. Spencer et al. in [11] presented a tool for sentiment analysis of twitter data named as sentimentor. It uses naïve bayes classifier to categorize tweets as positive negative or neutral. In [12], authors have extracted information from the tweets using keyword based knowledge extraction. The proposed approach enables the extraction of keywords entities, synonyms and parts of speech from tweets which were used for sentiment analysis. Bifet, A. et al. [13] in their other paper discussed the handling of tweets in real-time. They introduced a system that processes tweets in real time inspite of their dynamic nature and named it as MOA-TweetReader. This system has two basic tasks to perform: Firstly, it detects the changes in term frequencies and secondly, it performs sentiment analysis in real time. Go [14] proposed a new approach for automatically classifying twitter tweets as positive or negative using a query term. The main contribution of this paper is the idea of using tweets with emoticons for distant supervised learning such as :-) represents positive and :-(represents negative. Ye, S. and S. Felix Wu [15] found that twitter is a very widespread OSN where 700K users compose 58M messages. They proposed a technique to determine a message propagating pattern to show how breaking news spread all over the media. To evaluate the system they examined several social influences and their effects such as stabilities, correlations and assessments. In [16], the authors have presented an unsupervised system, SSA-UO. This system consists of three main phases: data pre-processing, contextual word polarity detection, and message classification. In pre-processing phase, spelling errors, emoticons, slang removal, lemmatization and POS tagging is performed.

3. Proposed Framework

The proposed system architecture consists of five main steps. The first step for sentiment analysis of twitter data is to acquire the data that has to undergo sentiment analysis. Raw tweets were obtained at the end of this stage and stored in a database which was the input for preprocessing step. Preprocessing consists of several steps which are the removal of URLs from the tweet, hashtag removal so that the tweet becomes more cleaned, removal of slangs, emoticon conversion to text, stop word removal. Filtered tweet is obtained at the end of this step which becomes the input of the Sentiment Influencers Identification step. Filtered text is passed to a tokenizer which separates each word in the tweet. This set of words (W) is passed to a POS tagger which tags the words of the tweet as nouns (N), verbs (V), adverbs (Adv) and adjectives (Adj). The output of this step is tagged text (T.T) which is a set containing noun, verbs, adjectives and adverbs. This set of POS is named as sentiment influencers. The output tagged text of the previous phase becomes the input of next step. The score of the T.T is calculated using SentiWordNet dictionary which assigns a score to each word in the tweet. The score obtained from SentiWordNet dictionary is named as standard score (S.S). Tagged score (T.S) is calculated by ranking the sentiment influencers which will be discussed in detail in implementation section. After the calculation of T.S, negation is checked in the tweet if it's present then the final score of the tweet is inversed that is multiplied with -1. The final score is calculated by summing the tagged score obtained by tagging the text and dividing it with the total number of words (W) in a tweet. The final score is passed to sentiment classifier which classifies the tweet into positive, negative or neutral. Sentiment classifier will classify the tweet as follows. If the final score of the tweet is greater than zero, the tweet is declared as positive. If the final score of the tweet is less than zero, the tweet is declared as negative. If the final score of the tweet is equal to zero, the tweet is declared as neutral.

Abstract level diagram of the proposed framework is presented in the figure 1 below. Detail of all the components is discussed in the next section.

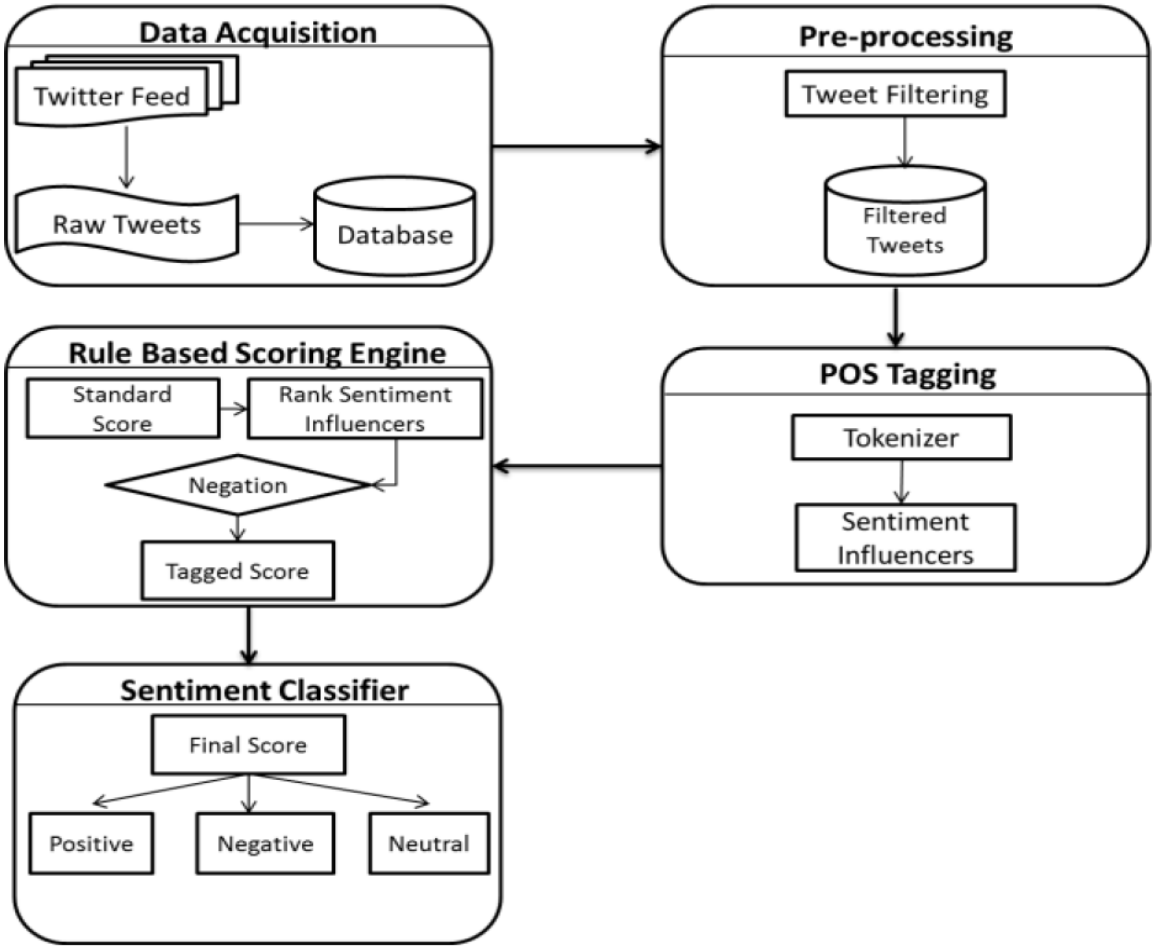

Figure 1. Abstract level diagram of Proposed Framework

4. Implementation

Complete framework has been implemented in Java language with Oracle database as backend data store. A number of APIs have also been used to supplement its functionality. Detail of each component is discussed under:

4.1 Data Acquisiton

The first step for sentiment analysis of twitter data is to gather the data that has to undergo sentiment analysis. For this purpose, twitter4J API [17] has been selected to acquire raw tweets. The twitter streaming API allows real time access to publicly available datAa on OSN. The query string was applied to obtain only English language tweets while discarding the rest. Raw tweets were obtained at the end of this stage and stored in a database which was the input for preprocessing step. Mathematically, set of tweets is represented as:

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

The diagrammatic representation of data gathering step is shown below in figure 2.

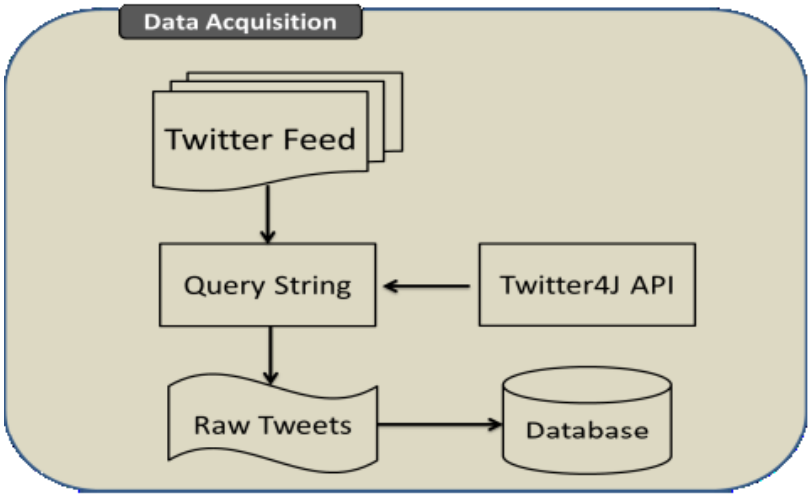

Figure 2. Data Acquisition

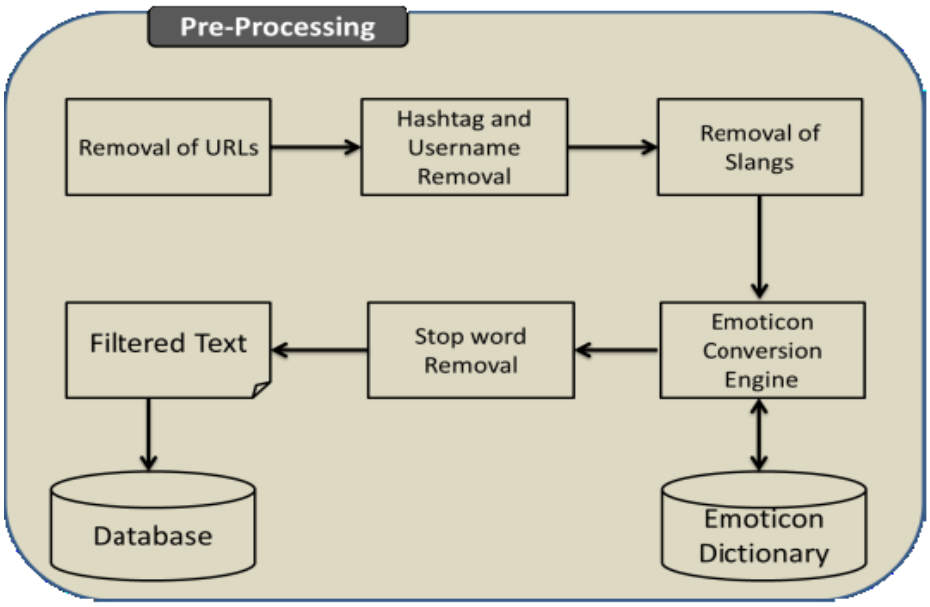

Figure 3. Pre-Processing

4.2 Pre-processing

Sentiment analysis of tweets encounters several issues such as poor spelling, use of abbreviations, poor punctuation, poor grammar, use of slangs. Therefore these hurdles need to be removed to ensure better results for sentiment analysis. Therefore, the proposed Pre-processing step consists of many subtasks. The first sub task in preprocessing is the removal of URLs from the tweet. For example the tweet before removing URL seems like this "*Photographer vs. White Sands | Brendan's Adventures http://buff.ly/1CQm19V #photography*". After removing the URL the tweet will be: "*Photographer vs. White Sands | Brendan's Adventures #photography*". The next subtask of pre-processing step is hashtag removal so that the tweet becomes more cleaned like this: "*Photographer vs. White Sands | Brendan's Adventures photography*".

The next step is the removal of slangs from the tweet. For this purpose, we have used the WordNet dictionary to check the meanings of the words whether they are proper words or slangs. The next step is to check for emoticons in the tweet. If any emoticon is present in the tweet, it will be replaced with the word such as :) will be replaced with 'Happy' and :(will be replaced with 'sad'. For this purpose, a simple emoticon dictionary is created, containing the most commonly used emoticons. If in a single tweet, more than one emoticon are found all of them will be converted to text and then scored accordingly. For example, "*Wifi makes me happy :)*" after converting the emotion the tweet will become "*Wifi makes me happy 'happy'*". Therefore, the score of happy will be doubled because of its occurrence. Emoticons play an important role in determining the polarity of a text if present. The next step of preprocessing is stop word removal. For this purpose, we have used textifier to check whether it is a stop word or not and then removed if identified as a stop word. After all these steps, the obtained tweets are filtered one and placed in a database for further processing. Figure 3 describes working and design of pre-processing component in detail.

4.3 Sentiment Influencers Identification

The input to this phase is the filtered text from the previous pre-processing phase. Filtered text is passed to a tokenizer which separates each word in the tweet. W is the set to total number of words in a tweet.

$$W = \{ w_1, w_2, ..., w_n \}$$

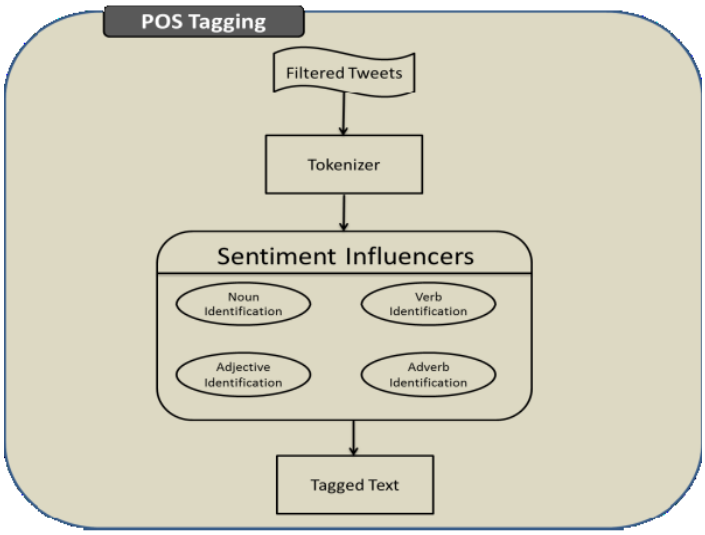

Figure 4. POS Tagging

A Part-of Speech (POS) tagger is a kind of software that reads some text in a particular language and then assigns parts of speech to each and every word in the text such as noun, verb, adverb, adjective etc. [46]. For this framework, the software that is used for POS tagging is called Stanford POS tagger. This set of words (W) is passed to a POS tagger which tags the words of the tweet as nouns (N), verbs (V), adverbs (Adv) and adjectives (Adj) also called sentiment influencers. A single tweet may contain more than one of it. The output of this step is tagged text (T.T) of sentiment influencers.

$$T.T = \{ N_{1...n}, V_{1...n}, Adj_{1...n}, ..., Adv_{1...n} \}$$

4.4 Rule Based Scoring Engine

The output tagged text of the previous phase becomes the input of this step. The score of the T.T is calculated using SentiWordNet dictionary which assigns a score to each word in the tweet. If the word is not found in the SentiWordNet dictionary, then it is

searched in Wordnet dictionary to check for its base words and synonyms and then that score is assigned to the word. The score obtained from SentiWordNet dictionary is named as standard score (S.S). Let's suppose v is the score of noun obtained from SentiWordNet dictionary in the tagged text, x is the score of verb, y is the score of adverb, z is the score of adjective in the tagged text (T.T) set. Then the Standard Score (S.S) is calculated as:

$$S.S = \{ v_{1...n}, x_{1...n}, y_{1...n}, ..., z_{1...n} \}$$

Next step is to calculate the tagged score. The T.T is checked for the presence of POS words if it is noun, the S.S for noun that is v will be multiplied with 1 because nouns are mostly neutral because they are used to describe some person, thing, place or idea. If the word is a verb then its score x is multiplied with 2 because verbs describe the actions of the nouns and plays a significant role in determining the polarity of a text for sentiment analysis. If the word is found to be an adverb, the score of the adverb y will be multiplied with 4 and given the highest weights because they modify verbs and adjectives and used as intensifiers. If the found word is an adjective, its score z will be multiplied with 3 in between the score of a verb and adverb because adjectives modify nouns. Mathematically, Tagged Score (T.S) is represented as:

$$T.S = \Sigma ((v_{1..n} * 1)(x_{1..n} * 2)(y_{1..n} * 3)(z_{1..n} * 4))$$

After the calculation of T.S, negation is checked in the tweet if it's present then the final score of the tweet is inversed that is multiplied with -1 which is represented as:

$$T.S = -1\left(\sum ((v_{1..n} * 1)(x_{1..n} * 2)(y_{1..n} * 3)(z_{1..n} * 4))\right)$$

4.5 Sentiment Classifier

Sentiment classifier takes the output from the last step that is the tagged score as input. The final score is calculated by summing the tagged score obtained by tagging the text and multiplying it with the respective number and dividing it with the total number of words (*W*) in a tweet. The mathematical representation of final score is given as below:

$$F.S = \frac{\Sigma (T.S)}{W}$$

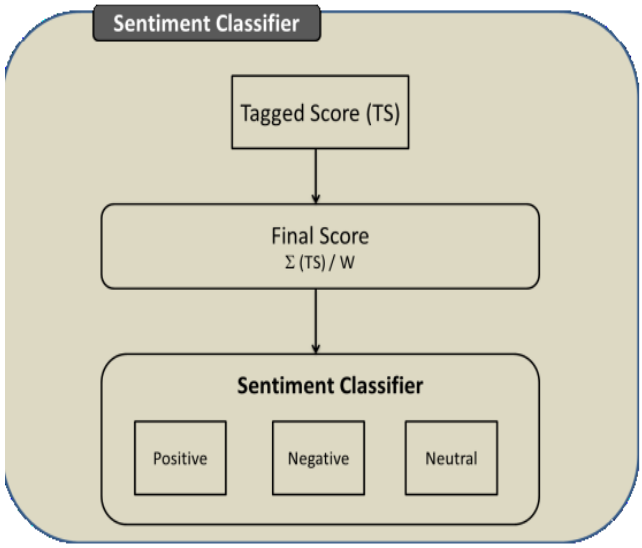

Figure 5. Rule Based Scoring Engine

The final score is passed to sentiment classifier which classifies the tweet into positive, negative or neutral. Sentiment classifier will classify the tweet as follows. If the final score of the tweet is positive then the tweet is declared as positive. Mathematically, it can be represented as:

$$Positive\ Tweet = F.S > 0$$

If the final score of the tweet is negative, the tweet is declared as negative.

$$Negative\ Tweet = F.S < 0$$

If the final score of the tweet is 0, the tweet is declared as neutral.

Neutral Tweet = F.S = 0

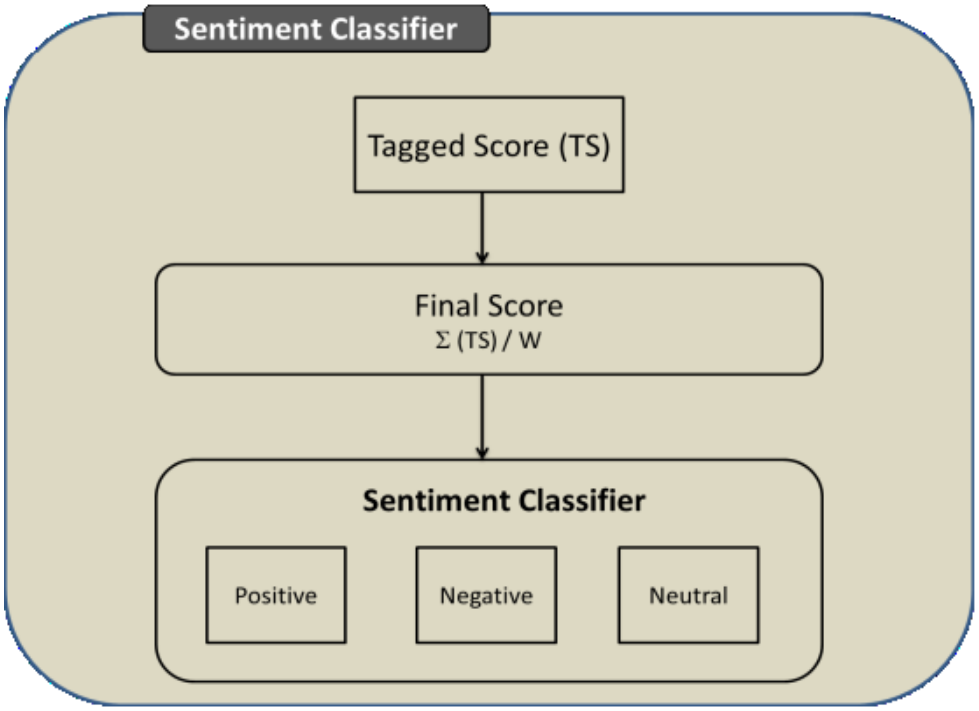

Figure 6. Sentiment Classifier

5. Results & Evaluation

This section evaluates results of proposed framework applied on STS-Gold dataset prepared from Stanford Twitter Sentiment Corpus. The dataset is publicly available and downloaded from [18]. The dataset contains a total of 498 tweets. Out of these 498 tweets, 182 are positive tweets, 177 are negative and remaining 139 are neutral tweets. After conducting experiment with proposed framework, satisfactory results were achieved. A confusion matrix for three classes (positive, negative and neutral) is shown in table 1.

	Predicted Positive	**Predicted Negative**	**Predicted Neutral**
STS Positive	152	17	33
STS Negative	23	149	5
STS Neutral	3	4	132

Table 1. Resuts

Most of the tweets have been rightly classified by proposed framework. The functionality of proposed framework can also be measured form precision and recall calculated from confusion matrix. The same is shown in table 3.

	Precision %	**Recall %**
Positive	85.39	83.52
Negative	87.65	84.18
Neutral	88	94.96

Table 2. Performance Measure

6. Conclusion & Future Work

Sentiment analysis is not a new term as significant amount of research has been carried out in this field. But there is always a space for improvement. Therefore this research has been carried out to improve the results using a novel unsupervised technique based on rule based scoring engine and ranking of sentiment influencers present in the tweet to categorize the tweet as positive, negative or neutral.

Future work to this research could be the use of supervised machine learning techniques along with sentiment influencers to improve the accuracy and results. A hybrid approach could be developed using various supervised machine learning techniques and then an unsupervised technique to develop better and more accurate results. Also, topic detection before performing sentiment analysis using any technique many also help in improving the results.

References

[1] Hu, M., Liu, B. (2004). Mining and summarizing customer reviews. *In*: Proceedings of the tenth ACM SIGKDD international conference on *Knowledge discovery and data mining*, August (p. 168-177). ACM.

[2] Melville, P., Gryc, W., Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. *In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge discovery and data mining*, June (p. 1275-1284). ACM.

[3] Pang., Lillian Lee., Shivakumar Vaithyanathan. (2002). Thumbs up? Sentiment classification using machine learning techniques. *In: Proceedings of the Conference on Empirical Methods in NLP*, p 79–86, Philadelphia, PA.

[4] http://webtrends.about.com/od/web20/a/social-media.htm

[5] http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/

[6] Turney, Peter. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of 40th Meeting of the Association for Computational Linguistics*, p 417–424, Philadelphia, PA.

[7] Benamara, F., Cesarano, C., Picariello, A., Recupero, D. R., Subrahmanian, V. S. (2007). Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone. *In:* ICWSM, March.

[8] Gimpel., Kevin., et al. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments, *In*: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: *Human Language Technologies: short papers*-2. *Association for Computational Linguistics*.

[9] Hatzivassiloglou, V., McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. *In*: Proc. 8th Conf. on European chapter of the Association for Computational Linguistics, Morristown, NJ, USA, *Association for Computational Linguistics* , 174-181.

[10] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. (2011). Sentiment analysis of twitter data. *In*: *Proceedings of the Workshop on Languages in Social Media* (p. 30-38). *Association for Computational Linguistics*, June

[11] Spencer, J., Uchyigit, G. (2012). Sentimentor: Sentiment analysis of twitter data. *In*: Proceedings of European Conference on *Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (p. 56-66).

[12] Batool, R., Khattak, A. M., Maqbool, J., & Lee, S. (2013, June). Precise tweet classification and sentiment analysis. *In*: *Computer and Information Science (ICIS), 2013 IEEE/ACIS 12th International Conference on* (p. 461-466). IEEE.

[13] Bifet, A., Holmes, G., Pfahringer, B. (2011). MOA-TweetReader: Real-Time Analysis in Twitter Streaming Data. *In*: T. Elomaa, J. Hollm´en, and H. Mannila (Eds.): DS 2011, LNCS 6926, Springer-Verlag Berlin Heidelberg, p 46-60.

[14] Go, A., Bhayani, R., Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1-12.

[15] Ye, S., Felix, S., Wu. (2010). Measuring Message Propagation and Social Influence on Twitter.com. *In*: Bolc, L., Makowski, M. Wierzbicki, A. (Eds.): SocInfo, LNCS 6430, p. 216–231, *Springer-Verlag Berlin Heidelberg*, p. 216-231.

[16] Ortega, R., Fonseca, A., Montoyo, A. (2013, June). SSA-UO: unsupervised Twitter sentiment analysis. *In*: *Second Joint*

Conference on Lexical and Computational Semantics (SEM)* (2, p. 501-507).

[17] Twitter4J, http://twitter4j.org/en/index.html Accessed 4 Feb 2013

[18] STS-Gold, http://www.sentiment140.com . Accessed 12 Mar 2015