Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SFT: A MODEL FOR SENTIMENT CLASSIFICATION USING SUPERVISED METHODS ON TWITTER

[1]RAZIEH ASGARNEZHAD, [2,*]S. AMIRHASSAN MONADJEMI, [3]MOHAMMADREZA SOLTANAGHAEI AND [4]AYOUB BAGHERI

[1,3]Department of Computer Engineering, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran

[2]Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran

[4]Department of Electrical and Computer Engineering, University of Kashan, Kashan, Iran

E-mail: [1]r.asgarnezhad@khuisf.ac.ir, [2]amonadjemi@yahoo.co.uk, [3] soltan@khuisf.ac.ir, [4]a.bagheri@kashanu.ac.ir

ABSTRACT

Twitter Sentiment Classification is one of the most popular fields in information retrieval and text mining. Thousand of millions of people around the world intensity use web sites such as Twitter. Twitter, as a micro-blogging system, allows users to publish tweets to tell others what they are thinking. In fact, there are already many web sites built on the Internet providing a Twitter sentiment search service. In those web sites, the user can input a sentiment target and in searching for tweets containing positive and negative sentiments. As a result of the increasing number of tweets over the past few years, tweets have attracted more and more attention. This is striking for consumers to research the sentiment of products before purchase automatically. This paper proposes a novel model for Twitter Sentiment Classification. The purpose of this model is investigating what is the role of weighting feature techniques in Sentiment Classification using supervised methods on the Twitter data set. Also, it explores binary classification which is classified data set into positive and negative classes. It is shown that usage of the proposed model can improve 7% the accuracy of Twitter Sentiment Classification. The results confirmed the superiority of the proposed model over the state-of-the-art systems.

Keywords: *Sentiment Classification, Support Vector Machine, Supervised Method, Twitter*

1. INTRODUCTION

The problem of automatic extraction of sentiment from informal text such as tweets is a recent area of investigation. Thousand of millions of people use Twitter extensively. Twitter allows users to publish tweets to tell others what they are doing or thinking. With increasing number of tweets over the Webs, tweets have attracted more and more attention. There is extensive interest in Sentiment Analysis of tweets with a variety of domains [1], [2], [12], [13], [14], [15].

There are many web sites on the Internet to provide a Twitter sentiment search service. In those web sites, the user can input a sentiment target and search for tweets including positive, negative, or neutral sentiments. This problem needs to adjust a query for classifying the sentiments of the tweets according to that query [3], [9]. Polarity analysis on micro-blogging is a very recent problem, so there are very few free resources and not available due to modified authorization status. However, we were unable to download all data sets.

Several researches in Twitter Sentiment Classification have focused on the usage of traditional classifiers and machine learning based classifiers, like Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machine (SVM) [3], [10]. These classifiers are trained on labeled corpora. Main problem in supervised learning techniques is the availability of trained data set. In contrast, unsupervised learning algorithms are a remarkable case so that they do not require any training data [14]. In addition, their goal depends only on situations and they are not mathematically well defined. We can only get a few numbers of data sets for supervised models because manually gathering of training sets is very time consuming. Authors in 2015 was compared the validity of supervised and unsupervised approaches [4].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Bagheri et al. proposed two unsupervised models for Sentiment Analysis of reviews [5]. However, tweets are usually more ambiguous than other sentiment data such as reviews and blogs.

Most of challenges can be considered in Twitter Sentiment Classification with respect to other data set s: classification accuracy, data sparsity problem, linguistic representational and tweets are very short and often show limited sentiment cues. In this study, we tried to determine which way can improve the model performance for Twitter. By considering the sentiment labels of the related tweets, we can further boost the performance of the model, especially for very short. These related words of tweets provide rich information about what the given tweet expresses and should be taken into consideration for classifying the sentiment of the given tweet.

In this paper, we propose a novel model, called SFT (short for supervised, feature, and twitter), to handle accuracy challenge and improve Twitter Sentiment Classification. Specifically, boosting method, weighting feature mechanisms like Term Frequency (TF) and Term Frequency–Inverse Document Frequency (TFIDF) weighting mechanisms were applied to incorporate features generated using words connected with the given target in the tweet. The proposed model is novel, because apply labeled binary data set and combine a set of different preprocessing steps for Twitter Sentiment Classification. We will reveal that a combination of different preprocessing and used classifiers play a major role in the performance of the model. In addition, we will show that the usage of supervised method and combined with weighting feature mechanisms can improve the accuracy of Twitter Sentiment Classification. To conduct the model, we will use Sanders have collected from Oct 15, 2011 to Oct 20, 2011 for Apple Corporation on four different topics (Apple, Google, Microsoft, and Twitter).

The rest of this paper is organized as follows: Section 2 shows a summary of the related works. The proposed model is presented in Section 3 and evaluated in Section 4. Finally, conclusion and future works present in Section 5.

2. RELATED WORKS

A large number of methods have been proposed for improving the accuracy of Twitter sentiment classification. Most of these methods found in literature are focused mainly on improving the performance of the classifier by changing the architecture of the classifier and manipulating the parameters for binary classifications. Taxonomy of methods was shown in Figure 1. The following text describes a brief review on some salient approaches in the literature.

One of the first studies on Twitter polarity classification was done in 2009 by Go et al. They introduced an approach for classifying the sentiment of Twitter messages which are classified as either positive or negative. They presented the results of machine learning algorithms for classifying the sentiment of Twitter messages using distant supervision. It has been showed that machine learning algorithms (NB, ME, and SVM) have accuracy above 80%. They also described the Pre-processing steps is vital in order to achieve high accuracy. They collected data from API (short for Application Programming Interface) by query term. Both unigrams and bigrams were used as features. They used POS (short for part of speech tags) as features because the same word may have many different meanings depending on its usage. They explored the usage of unigrams, bigrams, unigrams and bigrams, and POS as features. Finally, the accuracy obtained 83% [3].

Following the work of Go et al. (2009), Liu et al in 2009 proposed a novel model, namely ESLAM, to handle both manually labeled data and noisy labeled data for training. The main idea was to train a language model based on the manually labeled data, and then used the noisy emoticon data for smoothing. The publicly available Sanders Corpus was used for evaluation that consisted of 5513 manually labeled tweets for four different topics (Apple, Google, Microsoft, and Twitter). Experiments on real data sets demonstrate that ESLAM can effectively integrate both kinds of data to outperform those methods using only one of them [6].

Based on these previous works, Pak and Paroubek build a sentiment classifier via the collected corpus to determine positive, negative, and neutral sentiments for a document. For pre-processing, a collection of operations including filtering, tokenizing, removing stop words, and constructing n-grams were used. Their classifier was based on the multinomial NB classifier that used n-gram and POS as features. Experimental evaluations show that their proposed techniques were efficient and performed better than previously proposed methods [7].

Speriosu et al. compared three main approaches including lexicon based ratios, maximum entropy classification, and label propagation. They

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

evaluated these models on the Stanford Twitter Sentiment (STS), Obama-McCain Debate (OMD), and Health Care Reform (HCR). It was shown that a maximum entropy classifier trained with distant supervision works better than a lexicon based ratio predictor, improving the accuracy from 58.1% to 62.9%. By using the predictions of classifier in combination with a graph to incorporate tweets and lexical features, they obtain even better accuracy of 71.2% [8].

Jiang et al. focused on target dependent Twitter Sentiment Classification to classify the sentiments of the tweets as positive, negative, or neutral. By incorporating target dependent features and taking related tweets into consideration, Twitter Sentiment Classification was improved. A binary SVM classifier (SVM-Light with a linear kernel) is built to perform the classification. Pre-processing was containing tweet normalization, POS, word stemming, and syntactic parsing. They collected tweets using the Twitter API and manually classified each tweet as positive, negative, or neutral towards the query with which it is downloaded. According to the experimental results, the highest accuracy was 85.6% which belongs to target dependent features. They were found that target dependent features significantly outperform the previous target-independent classifiers [9]. In 2012, Saif et al. proposed two sets of features to alleviate the data sparseness problem. One is the semantic feature set where we extract semantically hidden concepts from tweets. Another is the sentiment feature set where we extract latent topics and the associated topic sentiment from tweets. Experimental results on the Stanford Twitter Sentiment Data set [1] showed that both feature sets outperform the baseline NB model using unigrams only. We have tested both the NB classifier from WEKA and the ME model from MALLET. Their results show that NB consistently outperforms ME. Moreover, using semantic features rivals the previously reported best result. Using sentiment topic features achieved 86.3% Sentiment Classification accuracy[10].

In 2013, Hu et al. proposed a novel sociological approach to handle networked texts in micro-blogging. They investigated whether social relations can help Sentiment Analysis by proposing a Sociological Approach to handling Noisy and short Texts (SANT) for Sentiment Classification. In particular, they presented a mathematical optimization formulation that incorporates the sentiment consistency and emotional contagion theories into the supervised learning process; and utilize sparse learning to tackle noisy texts in micro-blogging. They extracted sentiment relations between tweets based on social theories, and model the relations using graph Laplacian. Experimental results showed that the user-centric social relations are helpful for Sentiment Classification of micro-blogging messages. Empirical evaluations demonstrate that this model significantly outperforms the representative Sentiment Classification methods on two real-world data set s, and SANT achieved consistent performance for different sizes of training data, a useful feature for Sentiment Classification. The value of accuracy 79.6% and 76.3% obtained for STS and OMD, respectively [11].

Hassan khan et al. presented an algorithm for twitter classification based on a hybrid approach. Their method includes various Pre-processing steps before feeding the text to the classifier. The main goal of this research is to improve the accuracy of text classification and resolve the data sparsity issues. In this research, each tweet is pre-processed using pre-process procedure and then classification is performed at each refined tweet as defined in classification procedure. Finally, the output is generated in the form of positive, negative or neutral labeled tweets. Six data set s were generated using the data acquisition module. The experiments have been conducted using six different data set s and are performed using 2116 random tweets. The highest accuracy was 88.89% which belongs to data set 1 [2].

Montejo et al. presented a novel approach to Sentiment Polarity Classification in Twitter posts by extracting a vector of weighted nodes from the graph of WordNet. These weights are used in SentiWordNet to compute a final estimation of the polarity. Therefore, the method proposed a non-supervised solution that is domain-independent. They built a corpus of tweets written in English following the Go et al. (2009). Experiment consisted of evaluating a supervised approach, like SVM, using the well known vector space model to build the vector of features using TFIDF weighing. Stop words have not been removed. The SVM-Light software with the default configuration parameters (linear kernel) was used to compute support vectors and to evaluate them using a random leave-one-out strategy. From a total of 376,284 valid samples, 85,423 leave-one-out evaluations were computed, reporting a value of

[1] http://twittersentiment.appspot.com/

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

0.6429, 0.6147 and 0.6285 for precision, recall and F1, respectively [14].

In 2015, Vo et al. proposed a method consist of five main stages including: (i) given a tweet, all its words are first mapped into distributed representations, (ii) before the left and right contexts of a given target are extracted, (iii) the full tweet, the left, and right contexts and their lexicon-based alternatives, (vi) are used for feature extraction, and (v) the resulting features are used as input for sentiment classification. Their experiments showed that multiple embeddings, multiple pooling functions, and sentiment lexicons offer rich sources of feature information, which leads to significant improvement on accuracies [15].

Table 1 summarizes an evaluation comparison among some of sentiment classification tasks on the different Twitter data set s and compared them. As it can be seen, the value of accuracy varies from 67% to 88.89% and it depends on the used data set. In this study, sanders data set was used for binary classification and results were compared to results obtained by Liu et al. 2012.

3. PROPOSED MODEL

The proposed model explores binary classification which are classified the data set into positive and negative classes. This model investigates the effects of TF and TFIDF mechanisms using supervised technique like SVM on the Twitter data set. The name of SFT is defined for this model. The efficacy of the method has been tested on the two data set s and compared with proposed methods in [6]. Previous works mostly have special attention to supervised learning methods [1], [3], [10], or combination of methods [21]. Experimental Results were illustrated via two experiments on the data set and validated that the SFT outperforms the existing methods on the data set. Maximum accuracy and F1 obtained 89.76% and 87.34%, respectively. The SFT highlighted in five phases; (1) pre-processing steps (2) first weighting feature mechanism, (3) second weighting feature mechanism, (4) classification method, and (5) performance evaluation. The graphical representation of the SFT with details has been shown in Figure 2.

A set of operations for text pre-processing were used. In addition, TF and TFIDF weighting mechanisms and n-gram features were employed. A term n-gram is defined as a series of consecutive tokens of length n. Most of the existing works applied unigram as the basic feature. Unigram is showed as a collection of unique words in each document where each word is implied as a feature. In this model, values one, two, and three as n were used. Classification methods were ran on test data set and were processed each tweet. It classifies the tweets into positive and negative classes. In fact, the supervised method like SVM was used and 10-fold cross-validation scheme on the data set was adjusted. Cross-validation estimates the statistical performance usually on unseen data set. It is mainly used to estimate how accurately a model will perform in practice.

Two experiments for achieving the state-of-the-art results using R Studio were applied. The purpose of this study is investigating what is the role of n-grams in Sentiment Classification applications using supervised method. In the rest of this section, phases of the model will describe.

3.1 Pre-processing Steps

Firstly, characters and words and filter useless tokens were tokenized. Tweets were split into a sequence of tokens and filters tokens. Using stemming algorithms the root of words were found. Then, the stop words from text via eliminating every token which equals word from the stop word list were filtered. Consequently, n-gram features including unigrams, bigrams, and trigrams were used.

3.2 First Weighting Feature Mechanism

TF is a particular term presented up in a document. In this mechanism, each number in the vector represents word's frequency in the document. TF is the relative frequency of a word in the document. TF is defined as total occurrences of the word *t* in the document *d* divided by total number of the words occurring in the document *d*.

$$TF = F_{t,d} / F_d \quad (1)$$

3.3 Second Weighting Feature Mechanism

TFIDF feature weighing mechanism was used to created word vector. TFIDF is one of the most famous weighting mechanisms and consists of two ratings, regularity and inverse regularity of phrase. Inverse document frequency is investigated by splitting the amount of records. In 2008, TFIDF mechanism defined by manning et al. as,

$$TFIDF = TF * \log N / F_t \quad (2)$$

Where TF is the frequency of word *t* in document *d*, *N* in the number of documents in collection, and Ft is in the number of documents in collection that contain word *t*. Indeed, TFIDF mechanism avoids

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

assigning high scores to words that occur too often in the data set [17].

3.4 Classification Techniques

The machine learning techniques often have been used in Sentiment Classification tasks [3]. In addition, previous works have showed that supervised methods are more accurate than others for text classification tasks [1], [2], [12]. In this paper, supervised method like SVM base learner was used to receive the highest performance.

SVM is a machine learning technique depending on the statistical learning concept. SVM have been applied efficiently in text classification and in a large range of series handling programs. This classifier makes the greatest accuracy outcome in text classification applications. The key role of the training step is to select a maximum margin hyper plane which is represented by vector $\vec{w}$. It separates the document vectors in one class from the others. Thus, an optimization problem is constrained. In here, cj ϵ {1, −1} (corresponding to positive and negative classes) is the correct class of document dj so that the solution can be illustrated as,

$$\vec{w} = \sum j \alpha j c j \vec{d_j}, \alpha j \geq 0 \quad (3)$$

Also, the αj could be gained by solving a dual optimization problem (Lagrangian multipliers). The $\vec{d_j}$ as the αj is greater than zero and is called support vector [17].

Experiments showed that SVM is one of the best supervised classifiers. SVM on the training data provided was trained. SVM is a state-of-the-art learning algorithm proved to be effective on text categorization tasks and robust on large feature spaces. The linear kernel and the value for the parameter epsilon=1.0 for LibSVM (C-SVC) were chosen by cross-validation on the training data. All relevant parameters had been set optimal.

3.5 Performance Evaluations

This subsection introduces measures for evaluating text classification. Confusion matrices, precision, recall, F1 and accuracy are used for evaluation of the SFT and comparison with other techniques. In first, measures for classification were considered. In machine learning techniques, P and N are the number of positive and negative tuples. True positives (TP) refer to the positive tuples that were correctly labeled by the classifier. Let TP be the number of true positives. True negatives (TN) are the negative tuples that were correctly labeled by the classifier. Let TN be the number of true negatives. False positives (FP) are the negative tuples that were incorrectly labeled as positive. Let FP be the number of false positives. False negatives (FN) are the positive tuples that were mislabeled as negative. Let FN be the number of false negatives. To evaluating the performance of the present model, seven criteria namely accuracy, positive/negative precision, positive/negative recall, and positive/negative f-measure were employed [16]. The confusion matrix is defined in Table 2 for two classes based cases.

Accuracy and precision measure the percentage of correctly classified instances of the unseen data. Accuracy calculates the sum of actual tuples that are classified as true positive (TP) and the number of true negative (TN) relative to the total number of classified instances. Precision is a measure which can be stated as the percentage of tuples which have been labeled as positive and actual. Recall is a measure which refers to the percentage of tuples which have been labeled positive. Precision and recall are related metrics. In other words, accuracy is the proximity of classification outcomes to the actual values without considering the class labels. Conversely, precision is the measure of accuracy provided that the patient class has been predicted.

Table 2: The confusion matrix

Confusion matrix		Predict class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

In fact, it measures the percentage of the items that the system detected (i.e., the system labeled as positive) that are in fact positive (i.e., are positive according to the human gold labels). Whereas, recall measures the percentage of items actually were presented in the input that were correctly identified by the system. F-Measure (F1) combines precision and recall into a single measure [16]. In practice, we generally combine precision and recall into a single metric called the F1. F1 comes from a weighted harmonic mean of precision and recall. These measures can be computed as in Eq. (3) to (6):

$$Acc. = (TP+TN) / (P+N) \quad (3)$$

$$Pre. = (TP) / (TP+FP) \quad (4)$$

$$Rec. = (TP) / (P) \quad (5)$$

$$F1 = (2\ Pre.\ Rec.) / (Pre. + Rec.) \quad (6)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. EXPERIMENT RESULTS

The corpus of Twitter for training and testing the conducted experimental evaluations were applied. These corpora were prepared by Sanders Analytics has collected from Oct 15, 2011 to Oct 20, 2011 for Apple Corporation on four different topics namely; Apple, Google, Microsoft, and Twitter. The detailed information of two corpora is shown in Table 3. This data set consists of 479 tweets. There are 163 positive and 316 negative tweets in the given data set.

Table 3: A description of the used data set

# of instances	# of instances in classes	
	Positive	Negative
479	163	316

4.1 Experiments

This subsection will present the obtained results on the Twitter data set. The experiments aim at evaluating the efficacy of n-grams, before performing the supervised method. Several pre-processing steps were evaluated to improve the performance of method in terms of accuracy and F1. The accuracy and efficiency of the SFT has been tested on the Twitter data set containing tweets on four different topics.

Experiment I. The first experiment investigated the effect of TF mechanism and n-gram features on the evaluation metrics. The present researchers were showed that 10-fold cross validation is better and used in this experiment.

Experiment II. The second experiment investigated the effect of TFIDF mechanism and n-gram features on the evaluation metrics. In this experiment, 10-fold cross validation is used as well.

The SFT model trained and the obtained results are shown in Tables 4 and 5. Table 4 shows the confusion matrix, average of F1, and the obtained results from experiment I. Evaluation parameters and accuracy were calculated by the confusion matrix. Also, the highest results in each column of table have been marked as bolded text. In Table 4, the highest accuracy was highlighted 89.67%. Also, the highest average of F1 achieved 87.34%. These results were belong bigram features and TF mechanisms. Table 5 shows the confusion matrix, average of F1, and the obtained results from experiment II. The highest accuracy and average of F1 achieved 89.35% and 86.70%, respectively. As you seen, these results were achieved when TFIDF mechanism is used. Indeed, TFIDF provide a sufficient ability to capture the sentiment expression patterns. But, TF may not be useful for Sentiment Analysis. It can found that all highest metrics are obtained when bigram features were applied. In contrary to Tables, it seems that a combination of bigram features and TF mechanism can improve the accuracy and F1. It can found that SVM confuses when n-gram with higher levels used.

Figure 3 presents the evaluation metrics in both experiments. It can reveal that all highest metrics are above 92%. Figure 4 compares the obtained accuracy and average of F1for the SFT model on the data set. It may reveal that the highest accuracy and average of F1are 89.76% and 87.34% which belongs to TF mechanism for experiment I. It can be shown that the SFT make batter for classification according to these data sets. Nonetheless, this model outperforms the existing methods for Sentiment Classification tasks.

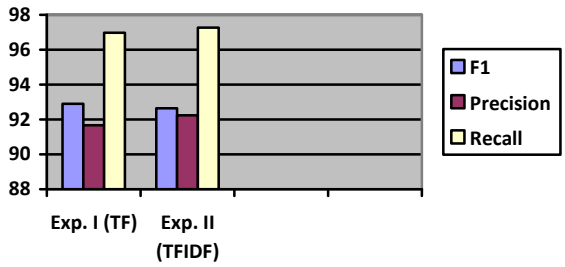

Figure 3: The average of evaluation metrics for the SFT model in term of weighting feature mechanisms

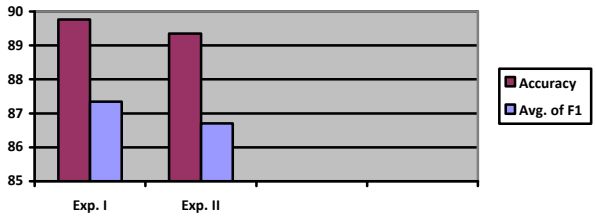

Figure 4: The comparison of accuracy value for the FST model

4.2 Comparison of empirical results

In this subsection, comparison of obtained results was shown. Table 6 was compared the performance among the SFT model, Liu et al. and other existing works in term of accuracy. It is clear from the comparison that the SFT shows better accuracy for classification task. It was found that there is a significant different between two weighting mechanisms. Figure 5 shows the comparison of accuracy value between the SFT and the ESLAM

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

model proposed by Liu et al in term of accuracy. For Twitter data set, the accuracy of the SFT obtained 89.76% using TF mechanism which is approximately 7% more than the value of accuracy that obtained by Liu et al. (Table 6). In the other hand, the accuracy of the SFT obtained 89.35% using TFIDF mechanism is approximately 7% more than the value of accuracy that obtained by Liu et al. [6]. The results indicated that the SFT is more accurate than its predecessors. According to two experiments, the used method yielded the best results. However, which method produces the better accuracy and whether the training data are complementary may depend on the type of features which is used (Figure 5).

Table 6: The comparison among the SFT model and others on different Twitter data set s in terms of accuracy

Data sets	*Acc.*	*Authors / Model*
Sanders	***89.76%***	*SFT (proposed)*
Sanders	*83%*	*Liu et al. (ESLAM)*
Twitter API	*83%*	*Go et al.*
Twitter API	*85.6%*	*Jiang et al.*
STS	*86.3%*	*Saif et al.*
STS-Gold	*79.65*	*Hu et al. (SANT)*
Stanford	*84.7%*	*Speriosu et al.*

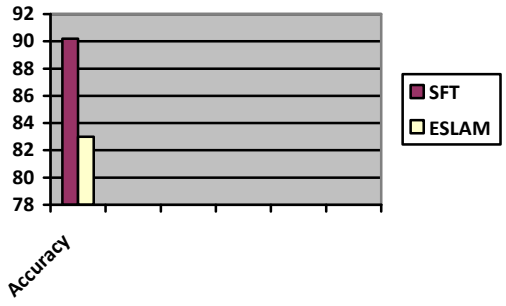

Figure 5: The comparison of accuracy value between the SFT and the ESLAM

5. CONCLUSION AN DFUTURE WORKS

The SFT proposed a model of Sentiment Classification for binary classifications which are classified the data set into positive and negative classes. Compared to the existing approaches on Twitter Sentiment Classification which either depend on sophisticated features or complicated learning procedure, the SFT are much more simple and honest. The effect of pre-processing steps using supervised method on the Twitter data set was investigated. Twitter is a micro-blogging system that allows users to publish tweets to tell others what they are doing or thinking. With growing the tweets on the Web, there are many fields to produce models for Sentiment Classification tasks. The highest average of F1 achieved 87.34% which belongs to TF mechanism. This due to the fact that TF mechanism is constructed the proper relationship of words to improved results. Experimental results illustrated that the present model outperforms the existing methods based on two experiments on the Twitter data set. Maximum accuracy obtained 89.76% which has improved approximately 7%. The SFT is very redeeming in comparison to others, since it used more related words using n-grams and supervised method.

The present researchers believe that the accuracy could still be improved. In future work, we will seek other approaches in Sentiment Classification to achieve higher accuracy. Also, we will investigate the model on other data sets. We could improve evaluation performance by 7% approximately. Although our results have been outperformed on Twitter data set, one of the most striking micro-blogging, we believe that our model is also useful for other data sets such as movie and product reviews.

REFRENCES:

[1] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!," Proceedings of the Fifth International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media, Barcelona, Spain , 2011, p. 538-541.

[2] F. H. Khan, S. Bashir, and U. Qamar, "TOM: Twitter opinion mining model using hybrid classification scheme," *Decision Support Systems,* vol. 57, 2014, pp. 245-257.

[3] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford,* vol. 1, 2009, p. 1-6.

[4] R. Asgarnezhad and K. Mohebbi, "A Comparative Classification of Approaches and Applications in Opinion Mining.", International Academic Institute for Science and Technology, Vol. 2, 2015, pp. 1-13.

[5] A. Bagheri, M. Saraee, and F. De Jong, "Care more about customers: unsupervised domain-independent aspect detection for

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

sentiment analysis of customer reviews," *Knowledge-Based Systems,* vol. 52, 2013, pp. 201-213.

[6] K.-L. Liu, W.-J. Li, and M. Guo, "Emoticon smoothed language models for twitter sentiment analysis," *Proceedings of* Aaai, 2012.

[7] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," Proceedings of LREC, Valletta, Malta, 2010.

[8] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge, "Twitter polarity classification with label propagation over lexical links and the follower graph," *Pr*oceedings of the First workshop on Unsupervised Learning in NLP, Edinburgh, Scotland, 2011, pp. 53-63.

[9] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human La*nguage Technologies-Volume 1, Portland, Oregon, 2011, pp. 151-160.

[10] H. Saif, Y. He, and H. Alani, "Alleviating data sparsity for twitter sentiment analysis," Proceedings of 2nd Workshop on Making Sense of Microposts (#MSM2012): Big things come in small packages at World Wide Web (WWW), Lyon, France, 2012, pp. 2-9.

[11] X. Hu, L. Tang, J. Tang, and H. Liu, "Exploiting social relations for sentiment analysis in microblogging," *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 537-546.

[12] H. Saif, M. Fernández, Y. He, and H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of twitter," Proceedings of 9th Language Resources and Evaluation Conference (LREC). Reykjavik, Iceland, 2014, pp. 810-817.

[13] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification," *Proceedings of ACL (1)*, 2014, pp. 1555-1565.

[14] A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López, "Ranked wordnet graph for sentiment polarity classification in twitter," *Computer Speech & Language,* vol. 28, 2014, pp. 93-107.

[15] D.-T. Vo and Y. Zhang, "Target-Dependent Twitter Sentiment Classification with Rich Automatic Features," *Proceedings of IJCAI*, 2015, pp. 1347-1353.

[16] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques: concepts and techniques*: Elsevier, 2011.

[17] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval* vol. 1: Cambridge university press Cambridge, 2008.

Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

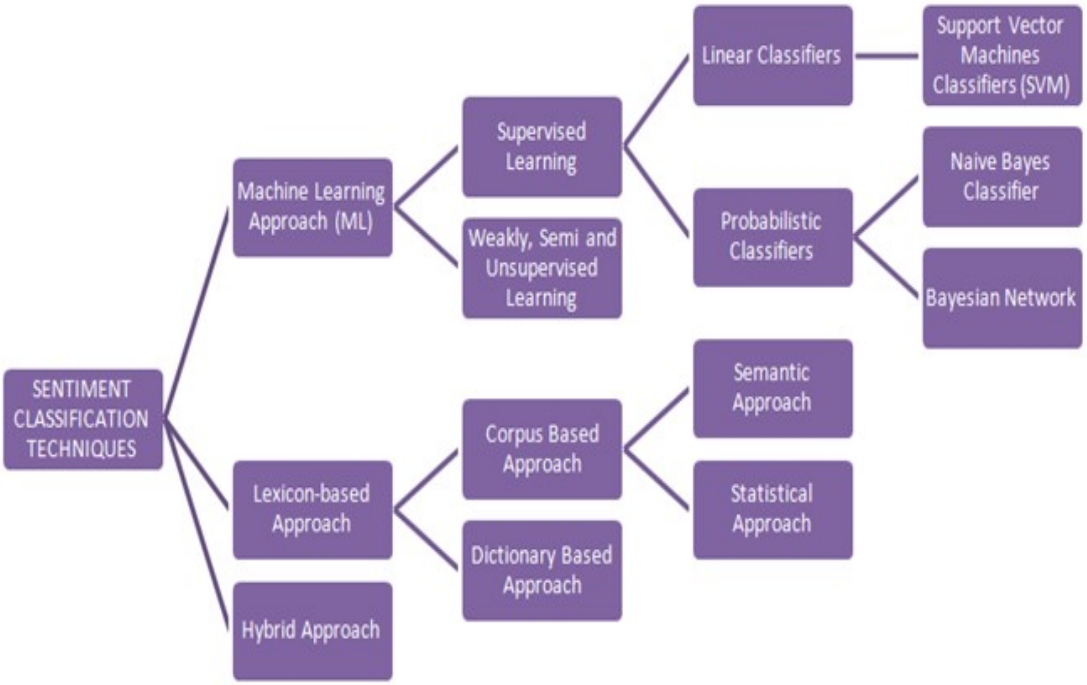

Figure 1 : Classification of opinion mining Approaches [4]

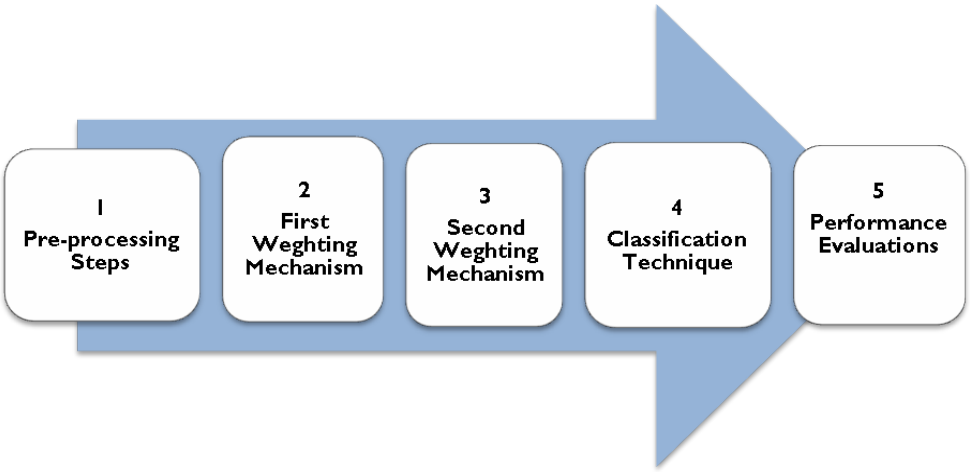

Figure 2: The steps of the SFT model

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: The comparative evaluation among existing tasks for Twitter sentiment

Authors	Used Approaches	Data set	Accuracy
Liu et al. 2009 (ESLAM)	integrate both manually labeled data and noisy labeled data, maximum likelihood estimate	Sanders	83%
Go et al. 2009	NB, ME, SVM, POS, n-grams	Twitter API	83%
Pak and Paroubek 2010	multinomial NB, SVM, part-of-speech, n-grams, statistical linguistic analysis	Twitter API (similar to Go et al., 2009)	above 81%
Speriosu et al. 2011	lexicon based ratios, ME, label propagation	Stanford HCR OMC	84.7% 71.2% 66.7%
Jiang et al. 2011	SVM, POS	Twitter API	85.6%
Saif et al 2012	Naive Bayes, ME, sentiment topic features, semantic features	STS	86.3%
Hu et al. 2013 (SANT)	graph Laplacian	STS-Gold OMC	79.6% 76.3%
Hassan khan et al. 2014	SentiWordNet	data set 1 data set 2 data set 3 data set 4 data set 5 data set 6	88.89% 82.86% 86% 85% 85.55% 85.90%
Montejo et al. 2014	WordNet, SentiWordNet, TF-IDF, Support Vector Machine	Go et al. (2009)	precision=64.29% recall =61.47% F1=62.85%
Vo et al. 2015	LibLinear	Dong et al. (2014)	69.1%

Table 4: The obtained results of the Experiment I (%)

N-gram	Actual / Predict	Confusion matrix		Results				
		Positive	Negative	Precision	Recall	F1	Accuracy	Avg. of F1
n=1	Positive	100	12	86.65	96.36	91.25	87.26	83.94
	Negative	49	318	89.29	67.11	76.63		
n=2	Positive	110	10	89.14	**96.97**	**92.89**	**89.76**	**87.34**
	Negative	39	320	**91.67**	73.83	81.79		
n=3	Positive	109	10	88.89	96.96	92.75	89.56	87.05
	Negative	40	320	91.60	73.15	81.34		

Table 5: The obtained results of the Experiment II (%)

N-gram	Actual / Predict	Confusion matrix		Results				
		Positive	Negative	Precision	Recall	F1	Accuracy	Avg. of F1
n=1	Positive	109	17	88.67	94.85	91.66	88.09	85.47
	Negative	40	313	86.51	73.15	79.27		
n=2	Positive	107	9	88.43	**97.27**	**92.64**	**89.35**	**86.70**
	Negative	42	321	**92.24**	71.81	80.75		
n=3	Positive	101	11	86.92	96.67	91.54	87.67	84.47
	Negative	48	319	90.18	67.79	77.40		