EFTSA: Evaluation Framework for Twitter Sentiment Analysis

Abdullah Alsaeedi*
Department of Computer Science, College of Computer Science and Engineering, Taibah University, Medina, KSA.

* Corresponding author. Email: aasaeedi@taibahu.edu.sa
Manuscript submitted Number 24, 2018; accepted January 12, 2019.
doi: 10.17706/jsw.14.1.24-35

Abstract: Sentiment analysis is a characteristic task that aims to detect the sentiment of opinions in content. Twitter sentiment analysis (TSA) is a promising field that has gained attention in the last decade. Investigators in the TSA field have faced difficulties comparing existing TSA techniques, as there is no agreed systematic framework. This means that the evaluation of existing techniques relies on selecting different datasets without meaningful justification. Another issue that arises when comparing different TSA techniques is that there are no unified metrics. Some researchers select classification accuracy and others choose recall, precision, and F-measure metrics. In this paper, we propose a framework called Evaluation Framework for Twitter Sentiment Analysis (EFTSA) for TSA evaluation based on individual or multiple datasets. This would help researchers compare their Twitter sentiment approaches against others.

Key words: Sentiment analysis, twitter, evaluation, metrics.

1. Introduction

Sentiment analysis is also called opinion mining and uses text mining, computational linguistics, and natural language processes to systematically determine, analyze, and examine opinions, emotions and subjective information. Sentiment mining focuses on deriving the sentiments of people sharing positive or negative comments, by analyzing a large number of text corpora [1]. Recently, sentiment analysis approaches have focused on examining opinions or feelings on different subjects, such as peoples' impressions regarding movies, product purchasing, and daily matters.

Social networks, such as Twitter, LinkedIn, and Facebook are considered elegant platforms that allow people to pose and express their opinions regarding life matters. Twitter is one of the most widespread social networking sites, as the number of active users is 330 million and the number of daily posted Tweets is 500 million according to the latest statistics in March 2018 [2]. Mining sentiment in social networking has gained attention in the last decade. Due to the nature of Twitter, posted Tweets have been mined and analyzed as part of marketing strategies [3] and reviewing customers reviews about products [4]-[6].

Twitter sentiment analysis (TSA) methods have focused on examining messages, named Tweets, to extract the sentiments or feelings expressed by the posted Tweets [7]. In the last decade, many techniques have been proposed for detecting sentiments in Twitter data. There are three categories of TSA techniques to detect the sentiment polarity of Tweets – lexicon-based approaches, machine learning approaches, and hybrid approaches. The sentiment polarity can be either positive, negative or neutral.

A number of challenges inhibit the process of detecting opinions on Twitter. Giachanou and Crestani [7] summarized a list of challenges faced by researchers who attempted to build an elegant TSA method. The length of Tweets is one of the biggest challenges, as they are limited to 140 characters. In addition, most TSA methods do not consider the relevance of topics when classifying Tweets based on their subject [7]. Besides the sparseness of Tweets, other factors, including misspellings and slang vocabulary, have a negative impact on the performance of a TSA [7].

Machine learning approaches rely on building a sentiment classifier to detect Tweets' encapsulating opinions and determine their sentiment polarities. These approaches can be classified into three groups: supervised, unsupervised, and ensemble techniques. Support vector machine (SVM), maximum entropy (ME), and naïve Bayes (NB) are supervised classifiers that are widely used to classify Tweets as positive, negative, and neutral. Machine learning classifiers can be effective, provided there is sufficient training data to train these classifiers on the selected features [7]. SVM classifiers are very effective in detecting the sentiment in Tweets [8]. One of the drawbacks of using these classifiers in detecting sentiment in Tweets is data sparseness [7], [8].

There are various TSA techniques that use classifiers in the previous literature. Anton and Andrey [9] used SVM and NB in their experiments and showed that SVM performed better than NB. In addition, SVM with unigram feature extraction obtained a precision accuracy of 81% and a recall accuracy of 74%. Go and Huang [10] employed NB, ME, and SVM to classify Tweets into categories. ME with both unigram and bigrams, was the best performing classifier and attained a classification accuracy of 83%. Malhar and Ram [11] used NB, SVM, ME, and Artificial neural network (ANN) classifiers to decide the sentiment polarity of Tweets. To reduce feature dimensionality, they [11] combined principal component analysis (PCA) with SVM and showed that this combination resulted in a classification accuracy of 92%. Pak and Paroubek [12] experimented with SVM, conditional random fields (CRF), multinomial NB (MNB) classifier, and various feature selection strategies. The best results were obtained using MNB with part of speech tags and n-grams features [12]. Anton and Andrey [13] showed that SVM with unigram features outperformed NB.

The lexicon-based TSA approaches rely on dictionaries to determine the polarity of Tweets. Lexicon-based approaches are useful, as no training data is needed. These approaches leverage dictionaries and lexicons annotated by pre-determined sentiment scores to identify the opinion score (polarity) of a given Tweet. There are many sentiment lexicons that can be used in detecting the sentiment of Tweets, such as SentiWordNet [14] and WordNet [15].

There are a number of lexicon -based TSA techniques in the literature. Hu, Tang, Gao, and Liu [16] introduced a framework called emotional signals for unsupervised sentiment analysis (ESSA). They modelled emotional signals using emotion correlation and emotion indication. The former is used to easily detect the sentiment polarity from posts and words. The latter is used to detect the correlation between posts and words. Azzouza, Akli-Astouati, Oussalah, and Bachir [17] proposed an architecture to determine sentiments, and to decide the polarity of Tweets. Their architecture relies on using a dictionary-based approach to identify the opinion polarity of Tweets. It consists of multiple modules. An acquisition module was used to collect Tweets by posing queries. An opinion analysis module was used to estimate the opinion score for emoticons and words, as well as their averages. Various experiments were conducted on the semantic evaluation of system challenge (SemEval) datasets to assess the performance of the proposed real-time architecture. The proposed system attained an accuracy score of 0.559 on the SemEval 2013 dataset, compared to the SSA-UO system proposed by Ortega, Fonseca, and Montoyo [18], which reached a score of 0.50. Paltoglou and Thelwall [19] proposed a dictionary-based approach to identify the sentiment polarity for the given text. Predictions was made by estimating the level of emotional intensity. Their lexicon-based methods attained higher F1 scores on various datasets, outperforming some supervised TSA

classifiers. Asghar, Khan, Ahmad, Qasim, and Khan [20] incorporated rule-based classifiers with an improved lexicon-based TSA, to minimize data sparseness issues and to enhance the classification accuracy of TSA models.

Approaches that use machine learning and lexicon-based techniques together are called hybrid approaches and aim to improve the sentiment detection. It is important to highlight that hybrid approaches gain the strengths of both machine learning and lexicon-based approaches. Filho and Pardo [21] proposed a hybrid TSA method, by combining three classifiers: supervised machine learning, rule-based, and lexicon-based. The experimental results showed that the hybrid system outperformed the individual TSA methods. The hybrid method attained an F-measure of 0.56, compared to 0.14, 0.448, and 0.49, obtained by the rule-based, lexicon-based, and SVM classifiers respectively. Ghiassi, Skinner, Zimbra [22] proposed a hybrid method that combined a dynamic artificial neural network (DAN2) sentiment analysis method and a reduced Twitter lexicon. The collected results exposed that the DAN2 method performed slightly better than the SVM classifier. Khan, Bashir, and Qamar [23] proposed a Twitter opinion mining (TOM) framework to mitigate the sparsity of Twitter data for sentiment classification. The framework consists of a SentiWordNet analysis, an emoticon analysis, and an enhanced polarity classifier. The experiments showed that the proposed framework attained an average harmonic mean of 83.3% on six various datasets. Asghar, Kundi, Ahmad, Khan, and Khan [24] incorporated four classifiers – a slang classifier, a general purpose sentiment classifier (GPSC), an emoticon classifier, and an enhanced domain specific classifier. The results showed that the proposed method in [24] attained an F-score of 0.87 compared to an F-score of 0.80 obtained by the TOM framework [23].

1.1. Research Problem

In the literature, there are various datasets that can be used to evaluate TSA methods. Datasets are selected randomly, and this selection is unjustified. Giachanou and Crestani [7] stated that there is lack of benchmark datasets which is one of the main issue in TSA domain. Besides, some TSA methods were evaluated using user-defined data set. This may bias the results, making comparisons with other methods that used different datasets very difficult or impossible.

There is no agreement on the metrics used to evaluate the performance of TSA, as some researchers use recall and precision while others use F-score only. In this paper, we establish new frameworks that can be used as a basis for evaluating TSA methods allowing researchers to compare new approaches with previous ones. In addition, we identify the metrics to be used for evaluating TSA approaches.

1.2. Motivation

The task of measuring machine learning TSA approaches is straightforward and well-established. It relies on a dataset to train and test models. Various datasets have been widely used for training and evaluating TSA methods. Text classification metrics have been used for evaluating the performance of TSA approaches. The reasoning behind this is to use test sets to measure the effectiveness of TSA methods in detecting sentiments.

This paper aims to establish an evaluation model framework that can be used for any further comparisons between various machine learning-based TSA techniques. The proposed framework will make the comparison task easy. Instead of evaluating such techniques using the user-suggested datasets, we aim to measure the performance of TSA techniques with unified-agreed datasets and evaluation metrics.

It is vital to tackle the challenges of comparing TSA approaches. Training and evaluating machine learning-based TSA techniques based on training and test sets that share the same domain will lead to higher bias scores [25]. Hence, it is important to measure the sentiment classifiers based on datasets obtained from diverse domains.

1.3. TSA Evaluation Metrics

In the text classification field, various measurements are widely used to assess the performance of text classification techniques. Table1 shows the confusion matrix used to present the evaluation metrics.

Table 1. Confusion Matrix for Evaluating the Performance of TSA Methods

Actual class	Predicted class as positive	Predicted class as negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

The simplest measure to evaluate a text classifier is classification accuracy. It is one of the most commonly used metrics to assess the performance of TSA in classifying Tweets.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is the ratio of Tweets that are correctly classified as positive, divided by the number of Tweets that are predicted as positive. This measures the exactness of the TSA methods.

$$Precision = \frac{TP}{TP + FP}$$

Recall is the ratio of positive Tweets that are classified as positive. It measures the completeness of TSA methods.

$$Recall = \frac{TP}{TP + FN}$$

The *F*-score is a mixture of both recall and precision. It is considered a suitable metric to assess the performance of TSA methods.

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

2. Twitter Sentiment Analysis Datasets

2.1. Stanford Twitter Sentiment Test Set (STS-Test)

The STS-Test dataset was presented by Go, Bhayani, and Huang [10] and consists of two distinct sets: training and test. 1.6 million Tweets were collected for the training set. The testing set was collected using queries posed in the Twitter API, and contained 182 positive, 177 negative and 139 neutrals Tweets. The drawback of the STS-Test dataset is that the test set is comparatively small. However, it is widely used in various TSA evaluation tasks. Various TSA methods used STS to evaluate their methods [27 ,26 ,10].

2.2. Sanders Dataset

The Sanders dataset, introduced by Niek Sanders, covers four topics - Microsoft, Apple, Google, and Twitter. It is made up of 5,512 Tweets, each marked as negative, positive, neutral, or irrelevant [28]. There are different ensemble TSA methods that have used the Sanders dataset for evaluating their performance

[29], [30].

2.3. SemEval-2013 Dataset (SemEval)

The SemEval-2013 dataset contains 5,810 positive, 2,407 negative, and 6,979 neutral Tweets. It was designed to assess Twitter sentiment methods in the semantic evaluation of system challenge (SemEval-2013). Mohammad, Kiritchenko, and Zhu [31] and Martínez-Cámara, Montejo-Ráez, Martín-Valdivia, and Ureña-López [32] used the SemEval-2013 dataset to evaluate their TSA methods.

2.4. Health Care Reform (HCR)

Speriosu, Sudan, Upadhyay, and Baldridge [26] created the HCR dataset using Tweets about health care reform in the USA. It consists of 541 positive, 1381 negative, and 470 neutral manually-annotated Tweets. The HCR dataset is extracted by crawling Tweets posted with the "#hcr" hashtag. Saif, He, and Alani [33] used the HCR dataset in their supervised machine learning Twitter sentiment classifiers. da Silva, Hruschka, and Hruschka [29] used HCR for assessing their ensemble method.

2.5. Obama-McCain Debate (OMD)

Shamma, Kennedy, and Churchill [34] searched three hashtags during the Obama-McCain debate to build the dataset. The gathered dataset is comprised of 3,238 Tweets, which were manually marked as negative, positive, or neutral. This dataset is used in lexicon-based methods such as [16]. Saif, He, and Alani [33] evaluated their supervised learning TSA method using the OMD dataset. da Silva, Hruschka, and Hruschka [29] also used the OMD dataset to evaluate their ensemble classifier.

2.6. Sentiment Strength Twitter Dataset (SS-Tweet)

Thelwall, Buckley, and Paltoglou [35] built a dataset named the Sentiment Strength Twitter Dataset, intending to assess the SentiStrength method for detecting the sentiment strength of Tweets. The dataset was comprised of 4242 Tweets that were manually labelled with the corresponding sentiment strengths. The negative sentiment strength ranged between -1, not negative, and -5, extremely negative. The sentiment strength of positive Tweets ranged from 1 to 5, denoting not positive and extremely positive, respectively. Instead of relying on sentiment strengths, Saif, Fernandez, He, and Alani [28] re-annotated this dataset with polarity labels, such as positive, negative, or neutral to more easily measure the subjectivity of Tweets. Thelwall, Buckley, and Paltoglou [35] and Gao and Sebastiani [36] used this dataset in their studies.

2.7. The Dialogue Earth Twitter Corpus

The Dialogue Earth Twitter corpus (DETC) is composed of three various datasets the WA, WB, and GASP. The GASP contains Tweets about gas prices and consist of 12770 Tweets. The WA and WB sets comprise Tweets about the weather and contain 4490 and 8850 Tweets, respectively. Each Tweet was manually labelled as negative, positive, neutral, or not related. Asiaee, Tepper, Banerjee, and Sapiro [37] used the DETC to assess their TSA methods.

2.8. The STS-Gold Dataset

STS-Gold dataset was created by Saif, Fernandez, He, and Alani [28] and is a subset of the STS dataset that are introduced in [10]. It contains 2034 Tweets, 632 of which are positive and 1402 are negative. The STS-Gold dataset also comprises 27 positive, 13 negative, and 18 neutral Tweets.

3. The Proposed Frameworks

This section provides two evaluation frameworks. The first one is used to measure a given TSA method based on each dataset. The second framework is used to measure multiple TSA methods using multiple

datasets.

3.1. Evaluation of TSA Methods Based on Individual Dataset

In this framework, multiple well-known datasets are given to the TSA method to assess their performance. Fig. 1 illustrates the basic framework for evaluating a given TSA method, using each dataset separately. For each dataset, metrics will be computed to measure how effective the TSA being evaluated is in detecting and classifying the sentiment of Tweets. Thus, the F-score, precision, and recall can be computed to measure the efficiency of the TSA method. The average scores can be computed as well. The reason for evaluating a TSA method with different datasets is to avoid biased results which might not reflect the real performance.

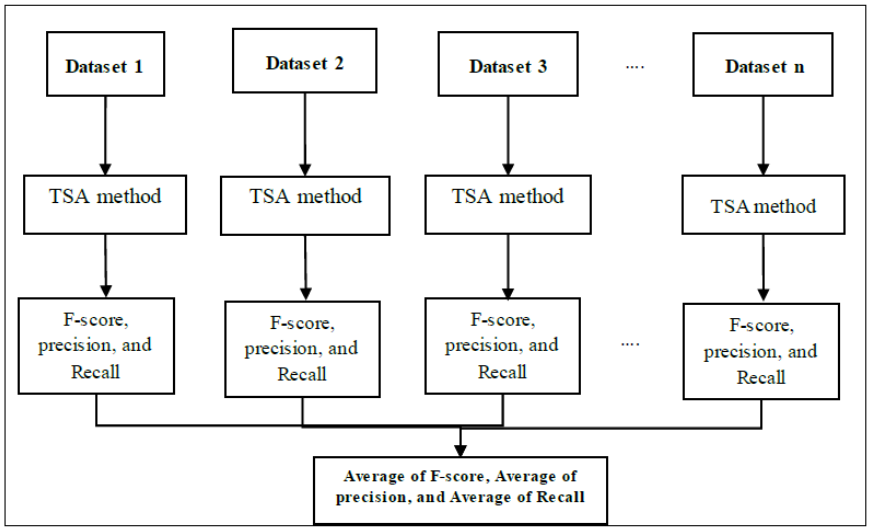

Fig. 1. Evaluation of a TSA method using individual datasets.

Input : $DataSets$, $TSAs$
/* $DataSets$ is a list of dataset and $TSAs$ is a list of twitter sentiment analysis method */
Result: $PerformanceScores$ Map $EvaluationMetrics$ scores to each TSA Method

$Datasets \leftarrow \{DS1, DS2, \cdots, DSn\}$;
$TSAs \leftarrow \{TSA1, TSA2, \cdots, TSAm\}$;
$PerformanceScores \leftarrow \{\}$;
for $D \in$ Datasets **do**
  $X, Y \leftarrow ReadDataSet(D)$;
  $Xtrain, Xtest, Ytrain, Ytest \leftarrow traintestsplit(X, Y)$;
  $Xtrain, Xtest \leftarrow processTweets(Xtrain, Xtest)$;
  $EvaluationMetrics \leftarrow \{\}$;
  for $t \in$ TSAs **do**
    $Classifier \leftarrow TSAmethod(Xtrain, Ytrain, t)$;
    $Predictions \leftarrow GetPrediction(Xtest, Classifier)$;
    $Metrics \leftarrow ComputeClassificationMetrics(Predictions, Xtest)$;
    $EvaluationMetrics \leftarrow \{EvaluationMetrics \cup (t, Metrics)\}$;
  end
  $PerformanceScores \leftarrow \{PerformanceScores \cup (D, EvaluationMetrics)\}$;
end
return PerformanceScores

Algorithm 1. Evaluation of TSA methods based on individual dataset.

The evaluation of TSA methods based on individual datasets is presented in Algorithm 1. The evaluation framework begins by providing different datasets, with TSA methods as shown in lines (1-2). The reason for providing these is to evaluate the performance of the TSA methods based on individual datasets. Then, for each dataset, the *ReadDataSet* function is used to read the Tweets, storing them in X and their target sentiment labels in Y. After that, the dataset is divided into training and testing datasets, as shown in line 6. The preprocessing stage is performed using the *processTweets* function. This includes stemming, lowercase conversion, URL symbol removal, @ symbol removal, # symbol removal, and stop word removal.
The iteration shown in lines 9-14 is intended to train classifiers, to obtain predictions, and compute the classification metrics iteratively for all TSA method and the dataset under consideration. The metrics includes classification accuracy, recall, precision, and the F-measure. Then, a pair (D, *EvaluationMetrics*) is added to the *PerformanceScore* map, to store the classification results for the current dataset under assessment, as illustrated in line 16. The process is iterative until all datasets are processed.

3.2. Evaluation of Various TSA Methods Based on Various Datasets

In this framework, all given datasets are given to different TSA methods to assess their efficiencies. These methods may belong to the same category, such as machine learning methods. This provides researchers with a framework to evaluate their methods compared to other methods proposed in the literature. Similar to the previous framework, the F-score, precision, and recall can be estimated to measure the efficiency of TSA methods.

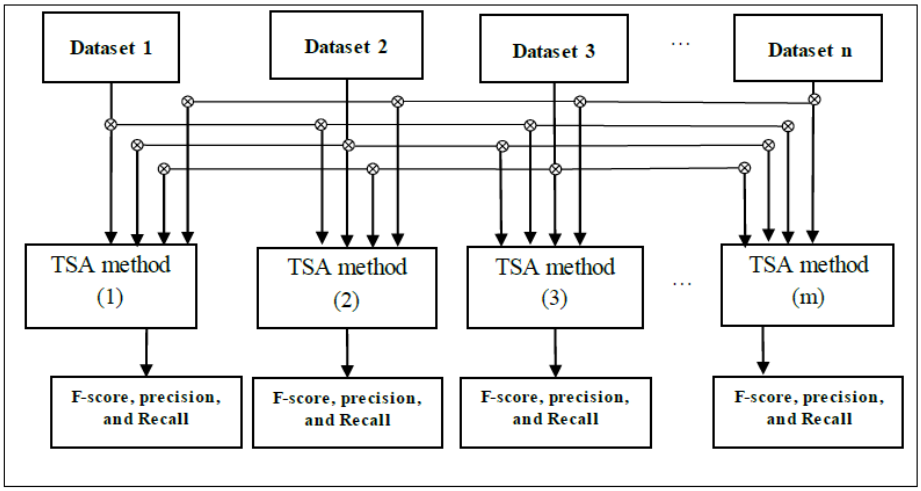

Fig. 2. Evaluation of multiple TSA methods using multiple datasets

The evaluation of multiple TSA methods based on multiple datasets is presented in Algorithm 2. The evaluation framework begins by providing different datasets and TSA methods, as shown in lines (1-2). The iteration in lines (5-7) read all datasets using the *ReadDataSet* function, and stores Tweets in X and their target sentiment labels in Y. The dataset is divided into training and testing sets as shown in line 8. The *processTweets* function is used to preprocess to the provided Tweets. The iteration shown in lines (11-16) train classifiers, obtain predictions after the training process, and compute the performance metrics. As we mentioned in Algorithm 1, the evaluation metrics include classification accuracy, recall, precision, and the F-measure. Then, the *EvaluationMetrics* map is updated after storing the performance metrics for the current TSA under assessment as illustrated in line 15. The process is iterative until all TSA are evaluated.

Input : $DataSets$, $TSAs$
/* $DataSets$ is a list of dataset and $TSAs$ is a list of twitter sentiment analysis method */
Result: $EvaluationMetrics$ Map $Metrics$ scores to each TSA Method

$Datasets \leftarrow \{DS1, DS2, \cdots, DSn\}$;
$TSAs \leftarrow \{TSA1, TSA2, \cdots, TSAm\}$;
$X \leftarrow \{\}$;
$Y \leftarrow \{\}$;
for $D \in$ Datasets **do**
  $X, Y \leftarrow \{X, Y \cup ReadDataSet(D)\}$;
end
$Xtrain, Xtest, Ytrain, Ytest \leftarrow traintestsplit(X, Y)$;
$Xtrain, Xtest \leftarrow processTweets(Xtrain, Xtest)$;
$EvaluationMetrics \leftarrow \{\}$;
for $t \in$ TSAs **do**
  $Classifier \leftarrow TSAmethod(Xtrain, Xtest, t)$;
  $Predictions \leftarrow GetPrediction(Xtest, Classifier)$;
  $Metrics \leftarrow ComputeClassificationMetrics(Predictions, Xtest)$;
  $EvaluationMetrics \leftarrow \{EvaluationMetrics \cup (t, Metrics)\}$;
end
return EvaluationMetrics

Algorithm 2: Evaluation of TSA methods based on multiple datasets.

4. Experimental Methodology

To prove the concept of the proposed evaluation frameworks, a number of datasets were selected, based on their popularity in the TSA field. The selected datasets were the HCR, Sanders, STS-Test, and SemEval-2013. Each dataset was split into training and test sets, as shown in the above-mentioned algorithms. It is important to highlight that 25% of the datasets were considered for testing and 75% for training the classifiers. The preprocessing stage included stemming, lowercase conversion, URL symbol removal, @ symbol removal, # symbol removal, and stop word removal. The unigram and bigram features were extracted using TfidfVectorizer, which belongs to the *sklearn* package in *python.*

Four classifiers were chosen for proving the concepts of the proposed frameworks. The selected classifiers were SVM, Bernoulli naïve Bayes, multinomial maïve Bayes, and linear regression. Precision, recall, and the F-measure were computed for each classifier and each dataset separately. After that, all four sets of training data were input into each classifier.

5. Experimental Results

Table 2 summarizes the evaluation results attained by the first framework comparing the four classifiers, using the HCR dataset only. It is obvious that there is no superior in the performance of the candidate classifiers. The Bernoulli naïve Bayes achieved the highest scores compared to other classifiers.

Table 2. Evaluation Results for HCR dataset

Classifier	Precision	Recall	F-measure
SVM	0.48	0.47	0.45
Bernoulli naïve Bayes	**0.5**	0.47	0.45
Multinomial naïve Bayes	0.49	0.47	0.44
Linear regression	0.48	0.47	0.45

Table 3 reports the evaluation results obtained by the four classifiers and the Sanders dataset only. This is

another illustration that proves the applicability of the first framework. There are no clear improvements in the performance of the classifiers. The Bernoulli naïve Bayes attained the highest F-scores compared to the other classifiers, while the multinomial naïve Bayes achieved the highest precision score among the classifiers.

Table 3. Evaluation Results for Sanders Dataset

Classifier	Precision	Recall	F-measure
SVM	0.78	0.80	0.73
Bernoulli naïve Bayes	0.79	0.80	**0.74**
Multinomial naïve Bayes	**0.83**	0.79	0.70
Linear regression	0.78	0.80	0.73

Table 4 summarizes the scores obtained after implementing the first framework on the four selected classifiers and the Stanford sentiment 140 dataset. There are no clear improvements in the performance of the classifiers, as the scores are almost the same.

Table 4. Evaluation Results for STS-Test Dataset

Classifier	Precision	Recall	F-measure
SVM	**0.79**	**0.79**	**0.79**
Bernoulli Naïve Bayes	0.78	0.78	0.78
Multinomial Naïve Bayes	0.78	0.78	0.78
Linear Regression	**0.79**	**0.79**	**0.79**

Table 5 reports the evaluation scores computed for four the classifiers using the SemEval-2013 dataset. It is apparent that there are slight improvements in the performance of the SVM and linear classifiers compared to other classifiers.

Table 5. Evaluation Results for SemEval-2013 Dataset

Classifier	Precision	Recall	F-measure
SVM	**0.71**	**0.68**	**0.65**
Bernoulli naïve Bayes	0.69	0.63	0.58
Multinomial naïve Bayes	0.66	0.65	0.62
Linear regression	**0.71**	**0.68**	**0.65**

Table 6 summarizes the scores attained by the classifiers using Algorithm 2, based on all datasets together. It is apparent that both the SVM and the linear regression performed slightly better than the other classifiers.

Table 6. Evaluation Results for all Datasets

Classifier	Precision	Recall	F-measure
SVM	**0.79**	**0.79**	**0.79**
Bernoulli naïve Bayes	0.78	0.78	0.77
Multinomial naïve Bayes	0.78	0.78	0.77
Linear regression	**0.79**	**0.79**	**0.79**

To summarize, it is clear that the performance of the classifiers is effective when the STS-Test is used to train them. This is due to the large number of Tweets in the training set. This agreed with the finding that the performance of classifiers is affected by the amount of training data.

6. Conclusions

This paper investigated the evaluation and comparison issue of Twitter sentiment analysis methods. Researchers in the TSA field evaluate their methods based on different datasets and different evaluation metrics. It is important to highlight that there was no systematic way to evaluate and compare two or more TSA methods. This paper established a systematic approach to comparison between various TSA methods. We introduced two frameworks that can be employed by researchers in the sentiment analysis field, especially for TSA methods. The benefit of our frameworks is that the comparison between different TSA methods, including supervised learning, ensemble, lexicon-based and hybrid, is much easier, as the same training data and the same evaluation metrics can be used.

References

[1] Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, *181(6)*, 1138-1152.

[2] Aslam. S. (2018). Twitter by the numbers: Stats, demographics & fun facts. Retrieved from: https://www.omnicoreagency.com/twitter-statistics/

[3] Kirtiş A. K., & Karahan, F. (2011). To be or not to be in social media arena as the most cost-efficient marketing strategy after the global recession. *Procedia - Social and Behavioral Sciences*, *24*.

[4] Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2018). Sentiment analysis on product reviews using machine learning techniques.

[5] Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., & Haruechaiyasak, C. (2012). Discovering consumer insight from Twitter via sentiment analysis. *J. UCS*, *18(8)*, 973-992.

[6] Vyas, V., & Uma, V. (2018). Approaches to sentiment analysis on product reviews. *Sentiment Analysis and Knowledge Discovery in Contemporary Business*.

[7] Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of Twitter sentiment analysis methods. *ACM Comput. Surv.*, *49(2)*, 1-41.

[8] Abirami, A., & Gayathri, V. (2016). A survey on sentiment analysis methods and approach. *Proceedings of the 2016 Eighth International Conference on Advanced Computing*.

[9] Liang, P.-W., & Dai, B.-R. (2013). Opinion mining on social media data. *Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management*.

[10] A. Go, Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report*.

[11] Anjaria, M., & Guddeti, R. M. R. (2014). Influence factor based opinion mining of Twitter data using supervised learning. *Proceedings of the 2014 Sixth International Conference on Communication Systems and Networks*.

[12] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining.

[13] Barhan, A., & Shakhomirov, A. (2012). Methods for sentiment analysis of Twitter messages. *Proceedings of the 12th Conference of FRUCT Association*.

[14] Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWwordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining..

[15] Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38(11)*, 39-41.

[16] Hu. X., Tang, J., Gao, H., & Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. *Proceedings of the 22nd International Conference on World Wide Web*.

[17] Azzouza, N., Akli-Astouati, K., Oussalah, A., & Bachir, S. A. (2017). A real-time Twitter sentiment analysis using an unsupervised method. *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics*.

[18] Ortega, R., Fonseca, A., & Montoyo, A. (2013). SSA-UO: Unsupervised twitter sentiment analysis.

Proceedings of the Second Joint Conference on Lexical and Computational Semantics.

[19] Paltoglou, G., & Thelwall, M. (2012). Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3(4)*, 66.

[20] Asghar, M. Z., Khan, A., Ahmad, S., Qasim, M., & Khan, I. A. (2017). Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. *PloS one*, *12(2)*.

[21] Filho, P. B., & Pardo, T. (2013). NILC_USP: A hybrid system for sentiment analysis in Twitter messages. *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*.

[22] Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, *40(16)*, 6266-6282.

[23] Khan, F. H., Bashir, S., & Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, *57*, 245-257.

[24] Asghar, M. Z., Kundi, F. M., Ahmad, S., Khan, A., & Khan, F. (2018). T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme. *Expert Systems*, *35(1)*.

[25] Maynard, D., & Bontcheva, K. (2016). Challenges of evaluating sentiment analysis tools on social media.

[26] Speriosu, M., Sudan, N., Upadhyay, S., & Baldridge, J. (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. *Proceedings of the First Workshop on Unsupervised Learning in NLP*.

[27] Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., & Varma, V. (2012). Mining sentiments from tweets. *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*.

[28] Saif, H., Fernandez, M., He, Y., & Alani, H. (2013). Evaluation datasets for Twitter sentiment analysis: A survey and a new dataset.

[29] Silva, N. F. F. D., Hruschka, E. R., & Hruschka, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, *66*, 170-179.

[30] Hassan, A., Abbasi, A., & Zeng, D. (2013). Twitter sentiment analysis: A bootstrap ensemble framework. *Proceedings of the 2013 International Conference on Social Computing*.

[31] Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets.

[32] Martínez-Cámara, E., Montejo-Ráez, A., Martín-Valdivia, M. T., & Ureña-López, L. A. (2013). Sinai: machine learning and emotion of the crowd for sentiment analysis in microblogs. *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*.

[33] Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. *Proceedings of the International Semantic Web Conference*.

[34] Shamma, D. A., Kennedy, L., & Churchill, E. F. (2009). Tweet the debates: Understanding community annotation of uncollected sources. *Proceedings of the First SIGMM Workshop on Social Media*.

[35] Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the Association for Information Science and Technology*, *63(1)*, 163-173.

[36] Gao, W., & Sebastiani, F. (2015). Tweet sentiment: From classification to quantification. *Proceedings of the 2015 IEEE/ACM International Conference on the Advances in Social Networks Analysis and Mining*.

[37] Tepper, A. A. T., Banerjee, M. A., & Sapiro, G. (2012). If you are happy and you know it... tweet. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management.*

Abdullah Alsaeedi received a B.Sc. in computer science from the College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia, in 2008, a M.Sc. in advanced software engineering from the University of Sheffield, Department of Computer Science, Sheffield, UK, in 2011, and a Ph.D. in computer science from the University Of Sheffield, UK, in 2016. He is currently an Assistant Professor and the Head of Computer Science Department at Taibah University, Madinah, Saudi Arabia. His research interests include software engineering, software model inference, grammar inference, machine learning, data mining and document processing.