Machine Learning and Semantic Orientation Ensemble Methods for Egyptian Telecom Tweets Sentiment Analysis

Amira Shoukry* and Ahmed Rafea

Department of Computer Science and Engineering, The American University in Cairo (AUC), Cairo, Egypt
E-mail: am_magdy@aucegypt.edu; rafea@aucegypt.edu
**Corresponding Author*

Received 17 December 2019; Accepted 14 April 2020;
Publication 03 June 2020

Abstract

The vast amount of data currently available online attracted many parties to analyze sentiments expressed in these data extracting valuable knowledge. Many approaches have been proposed to classify the posted content utilizing a single classifier. However, it has been proven that ensemble learning and combining multiple classifiers may enhance classification performance. The aim of this study is to improve the Egyptian sentiment classification by combining different classification algorithms. First, we investigated the benefit of combining multiple SO classifiers using different subsets from SATALex Egyptian lexicon. Second, we investigated the benefit of combining three classification algorithms; Naïve Bayes, Maximum Entropy and Support Vector Machines, adopted as base-classifiers. The experimental results show that combining classifiers can effectively improve the accuracy of Egyptian dataset sentiment classification. However, building these ensembles require more time for processing than the individual classifiers. The time needed depends on the number of classifiers used and the combination method used

Journal of Web Engineering, Vol. 19_2, 195–214.
doi: 10.13052/jwe1540-9589.1924

to combine these classifiers. Thus, the more classifiers used, the more time needed.

Keywords: Arabic sentiment analysis, lexicon based sentiment analysis, egyptian dialect, arabic opinion mining, ensemble learning.

1 Introduction

With the widespread usage of Facebook, Twitter, and other social networking websites, together with the great opinionated web contents coming from blogs and forums, sentiment analysis or opinion mining received considerable attention. Sentiment analysis is the task of identifying the semantic orientation of a piece of text as either positive, negative or neutral. In other words, sentiment analysis focuses on determining the overall tonality of a document or the attitude of a writer regarding the specified topic (Abbasi et al., 2008). Sentiment analysis can be carried out at many granularity levels: expression or phrase level, sentence level, and document level (Medhat et al., 2014). This study is interested in sentiment analysis for the Egyptian Arabic dialect language at the sentence level classifying whether a review, or a tweet, as holding an overall positive, negative or neutral sentiment.

There are different approaches for sentiment analysis: Machine Learning, lexicon-based or semantic orientation (SO) and hybrid approaches. The ML approach is a supervised approach in which a set of data labeled with its class such as "positive" or "negative" are converted into feature vectors to train a classifier, employing one of the ML algorithms, to infer that a combination of specific features yields a specific class. On the other hand, the SO approach is an unsupervised approach in which a sentiment lexicon is used to extract the sentiment terms composing the text, then combine them to produce an overall sentiment score for the whole text. Finally, the hybrid approach is a combination of the two precedent approaches benefiting from advantages of each approach. However, the performance of each approach depends on the features extracted for the language and domain of application.

Lately, ensemble learning in the area of sentiment analysis has been of a growing interest aiming at improving the overall prediction accuracy. Ensemble learning is the combination of multiple classifiers to obtain a more accurate and reliable classification in comparison with the single classifier (Zhou, 2012). Thus, studying the effectiveness of both individual supervised, unsupervised classifiers and ensemble methods for each dialect is of great importance.

In this work, the main research objective was to investigate to what extent using an ensemble of ML algorithms, and ensemble of SO lexicons could improve the performance of classification of Egyptian dialect tweets. This led to the following research questions:

1. Would using such ML ensemble improve the performance with a statistically significant difference when compared to using a single ML classifier?
2. Would using such SO ensemble improve the performance with a statistically significant difference when compared to using a single comprehensive sentiment lexicon?

The remaining of the paper presents our achieved work in building ML and SO ensembles for the Egyptian Arabic telecom tweets. Section 2 summaries the related work done in this area, while section 3 explains the process of developing the ML and SO ensembles. Section 4 describes the experiments conducted to evaluate the performance of each of these ensembles. Finally, Section 5 talks about the challenges, conclusion and future work.

2 Related Work

The ensemble methodology is the process of combining multiple models (classifiers). Each of these models solves the same original task, to produce a better composite model having more reliable and accurate decisions or predictions compared to using a single model. The idea of building a predictive model by integrating multiple models has been under investigation for a long time. Most of the work done in this area focuses on combining the deep learning and the Machine Learning (ML) techniques to produce a unified model, with very little work concerned with the semantic orientation (SO) approaches. In this section, we will present some of the systems which used the ensemble methodology to solve their classification problem for both ML and SO techniques.

Starting with the ML technique, (Catal and Nangir, 2017) presented a novel sentiment classification technique based on Vote ensemble classifier. They have utilized three individual classifiers: Bagging (SVM), Naïve Bayes, and SVM using the CVParameterSelection parameter optimization algorithm of Weka, for their Turkish sentiment classification problem. Their proposed approach achieved better performance than Naïve Bayes, which was reported the best individual classifier for used datasets, and Support Vector Machines.

Three Turkish sentiment datasets (Books, Movies, Shopping) were tested, and their best accuracies produced were: 86.13%, 83.68%, and 79.96%.

Likewise, (Oussous et al., 2018) investigated the performance and the efficiency of the ensemble method on the Arabic sentiment analysis specifically on the Moroccan reviews. They built a new Moroccan Arabic dataset consisting of 2000 tweets/comments, with a good balance between negative and positive sentiments. The ensemble method was applied for more accuracy by integrating three classification algorithms: NB, ME and SVM. The results showed that ensemble of classification algorithms performed better than the three individual classifier. Two Moroccan datasets were used (their built dataset (MSAC), SemEval) and their best f-scores were: 83.4%, and 84.2%.

On the other hand, for the SO technique, (Ohana et al., 2011) conducted a comparative study of lexicon based sentiment classification on multiple domains. They have showed that classification performance varies with the chosen lexicon and the domain it is applied to. Also, lexicons showed a tendency to perform better on either positive or negative documents while underperforming on the other category. In addition, they proposed an approach that combines the predictions of the different classifiers using the sum of all scores as the predictor forming a classifier ensemble. This classifier is further extended by introducing a score adjustment factor based on a term's relative frequency of occurrence extracted from a corpus. They tested on six datasets containing user generated reviews from different domains, and their highest accuracy was 80.79%, and the highest recall is 66.39%.

Moreover, (Augustyniak et al., 2014) introduced a new approach for lexicon extraction which was used to extract more than 15 lexicons. These extracted lexicons were used for sentiment polarity assignment of the review text. These sentiment polarity scores were combined in a sentiment polarity matrix to train a strong classifier, such as the decision tree method, to predict the sentiment of new documents. They showed that the combined lexicons with trained classifier are much faster than a supervised approach to sentiment classification while yielding similar accuracy. Their method also proved to be efficient and fast across all examined datasets. They used five datasets, and their best f-score was 50%.

Finally, (Augustyniaket et al., 2016) presented a lexicon-based ensemble approach to sentiment analysis that outperforms the lexicon-based method. The method consisted of two steps. First, they employed their own technique (called frequentiment) for automatic generation of sentiment lexicons and some publicly available lexicons. Secondly, an ensemble classification (fusion classifier) uses these predictions from the previous step as input, then

it combines these predictions into an overall prediction. They tried couple of classifiers such as Decision Tree, Extra Tree Classifier, and AdaBoost. They conducted comprehensive analysis based on 10 Amazon review data sets, and their highest F-score was 56.9% using RFT classifier.

3 Methodology

Ensemble method is a technique that combines several base models in order to produce one optimal predictive model. In other words, Ensemble Learning is a process in which multiple classifiers are strategically constructed to solve a particular problem. It combines a diverse set of learners (individual models) to improvise on the stability and the predictive power of the model. Also, we studied the way in which the results of each classifier are combined. There are several methods for result combinations like; (1) Averaging where the results of all the classifiers are added together and the average is calculated; (2) Maximum where only the maximum result produced by any of the classifiers is considered; (3) Majority voting where most common class label prediction produced by the classifiers is the one to be considered.

The main advantage of the ensemble methodology is to aggregate the results of all the selected models; thus, reducing the probability of selecting unsuitable or a wrong single classification model for a dataset.

The main goal of this work is to explore the effect of ensemble learning for both SO and ML approaches for use in the sentiment analysis tasks of the Arabic Egyptian tweets for the telecom community. To accomplish this goal, our research work has been targeting two main areas: (1) generation of different SO ensembles to measure their performance against a single SO classifier; (2) generation of different ML ensembles to measure the performance against a single ML classifier. Each of these areas is detailed in the following subsections.

3.1 SO ensemble method

The main idea in the SO ensemble learning is to build different classifiers each using different lexicon then combing the results of these classifiers to produce the final result. Given the fact that the lexical sentiment is greatly affected by the context, domain-specific sentiment lexicons are considered an important factor for computational social science (CSS) (Hamilton et al., 2016). With the help of domain-specific lexicons, social sentiment analysis considers factors such as genre, demographic variation,

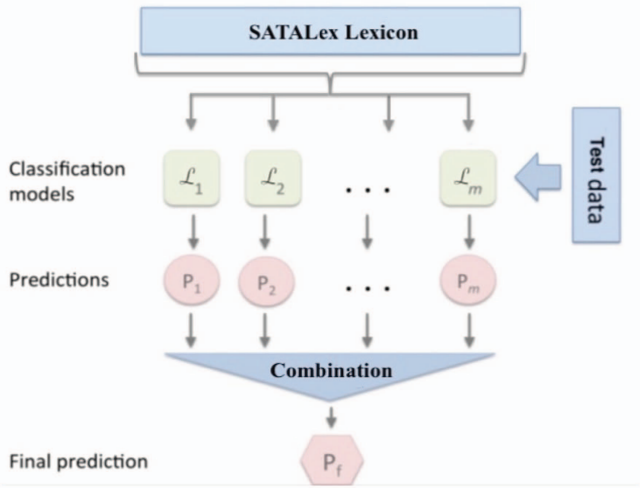

Figure 1 SO Ensemble Structure.

or community-specific dialect (Deng et al., 2014; Hovy, 2015; Yang and Eisenstein, 2015), without being biased towards domain-general contexts.

In this study, the process of building the subset lexicons depended mainly on the Egyptian semantic lexicon SATALex built in (Shoukry and Rafea, 2019). Different SO ensembles were created using two different parameters: (1) the number of subset lexicons; (2) the percentage of terms each subset lexicon represents from the main lexicon as shown in Figure 1.

Six sets of subsets lexicons were built: (1) 4 subset lexicons representing 50% of the main lexicon; (2) 4 subset lexicons representing 75% of the main lexicon; (3) 4 subset lexicons representing 90% of the main lexicon; (4) 8 subset lexicons representing 50% of the main lexicon; (5) 8 subset lexicons representing 75% of the main lexicon; (6) 8 subset lexicons representing 90% of the main lexicon. Moreover, since each lexicon produces a score for each tweet, we have applied three voting mechanisms (Average, Maximum, Majority). It is important to note that in each subset lexicon the terms were unique and not replicated.

3.2 ML ensemble method

The main idea in the ML ensemble learning is to combine multiple ML classifiers using decision formula to produce the final result. In other words, constructing and combining a set of hypotheses from the training data.

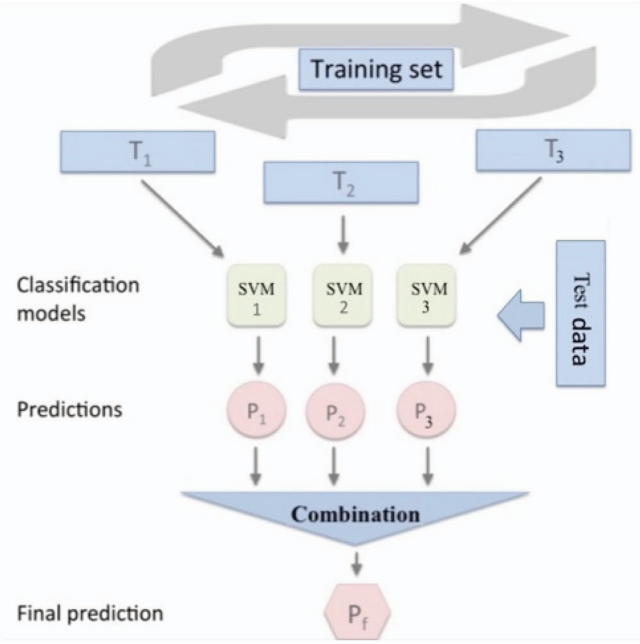

Figure 2 ML Ensemble 1 Structure.

Majority Voting was used to combine predictions from various classifiers. Hence, the final prediction and the class of the tweet is determined using the majority vote rule for the single votes produced by each classifier.

In this study, two ML ensembles were created: (1) combining three ML classifiers (SVM, NB, ME) built using the same training dataset; and (2) combining three SVM classifiers built using three different learning datasets generated from the main training dataset as shown in Figures 2 and 3.

For the first ML ensemble, we have followed the ML ensemble presented in (Oussous et al., 2018) for Arabic sentiment classification using the same three classification algorithms named Support Vector Machines (SVM), Naïve Bayes (NB) and Maximum Entropy (ME). The same training data was applied to each ML classifier. For the second ML ensemble, we have started by generating three different training datasets. The generation process involved selecting random tweets with replacement from the original dataset. Meaning that it is possible that some training tweets to be replicated or repeated in the generated training datasets. These generated training datasets had the same size as the original training dataset.

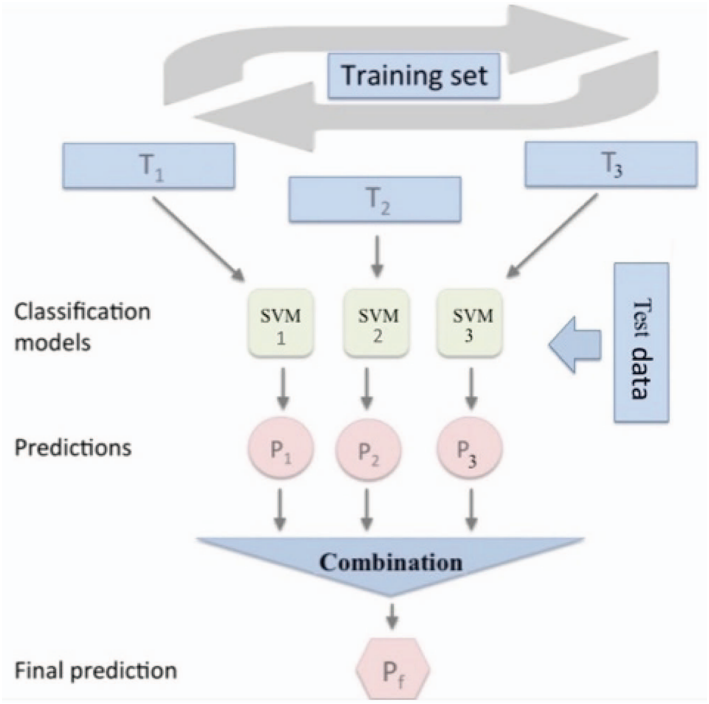

Figure 3 ML Ensemble 2 Structure.

Two main experiments were carried out for each ML ensemble. For both experiments, the tweets were represented using two methods as suggested by (Shoukry and Rafea, 2019) in representing the tweet vector: (1) the bag of words model using unigram presence; (2) hybrid model using sentiment words in the SATALex lexicon as features. In the first method, the feature vector for each tweet is represented as shown: (word1:0, word2: 1, word3: 0 . . . , "polarity"). Whereas the second method, the feature vector for each tweet is represented as shown: (senti_word1:0, senti_word2: 1, . . . , "polarity").

4 Evaluation

Following our proposed methodologies, we have carried out different experiments to compare their performance and discuss the results obtained. In this section, we present the details of the domain specific lexicon used; the datasets used and their distributions; and finally, the experiments conducted with their results.

4.1 The domain specific lexicons

The SATALex Egyptian semantic lexicon consists of a total of 1322 unique terms (94 positive single terms, 24 compound positives, 940 negative single terms, 264 compound negative).The main advantage of SATALex is that it includes some English transliterations terms like نيس (nice), اوفر (over), etc.. . . since they are commonly used in social media telecom domains. Although the subset lexicons were generated randomly from this main lexicon, it was clear that the negative terms and negative compound terms were more dominant in most of the generated subset lexicons than the positive terms and positive compound terms. This is mainly caused by the unequal distribution of the sentiment terms in the SATALex lexicon.

4.2 The used datasets

The Egyptian Company named RDI[1] appreciatively collected all the datasets to be used for research purposes with their main focus on the telecom domain. The annotation rules were set to annotate the collected datasets with annotation revisions to check and fix any erroneous annotation that could have taken place. The Egyptian train dataset consists of 8101 labeled tweets: 183 positive, 2597 negative, and 5321 neutrals. The Egyptian test dataset consist of 2692 labeled tweets: 77 positive, 943 negative, and 1672 neutrals. Given the unstructured nature of the used datasets, we have followed the approach proposed in (Shoukry and Rafea, 2012) for preprocessing, except for stemmer. So, only normalization and stop words removal were applied for preprocessing. Although the built train datasets were generated randomly by selecting random indexes with replacements from the main training dataset, it was obvious that the negative and neutral tweets were more dominant than the positive tweets. This dominance is caused by the great imbalance of the training dataset as people usually complain or criticize on social media more than they praise or compliment.

4.3 Experiments and results

The built subset lexicons and different learning datasets were used in two main experiments. The first experiment was to evaluate the performance of the proposed SO ensemble learning. While, the second experiment was to evaluate the performance of the ML ensemble learning. Finally, we compare the performance of ML ensemble learning against the SO ensemble learning.

[1] http://www.rdi-eg.com/

4.3.1 SO ensemble learning

Based on the methodology discussed in section 3.1, we wanted to evaluate the performance of the SO ensemble learning using the generated subset lexicons. All the three voting methods (averaging, maximum, majority) were applied.

Figures 4 and 5 show the F-score obtained after running the SO classifier using the 6 sets of subset lexicons: using 4 subset lexicons and using 8 subset lexicons with the three voting methods. From the first graph, the highest f-score (78%) was produced at: (1) lexicon size 75% using average and maximum voting methods; (2) lexicon size 90% using Majority voting method. While from the second graph, the highest f-score results (78%) was produced at: (1) lexicon size 50% using average and maximum voting

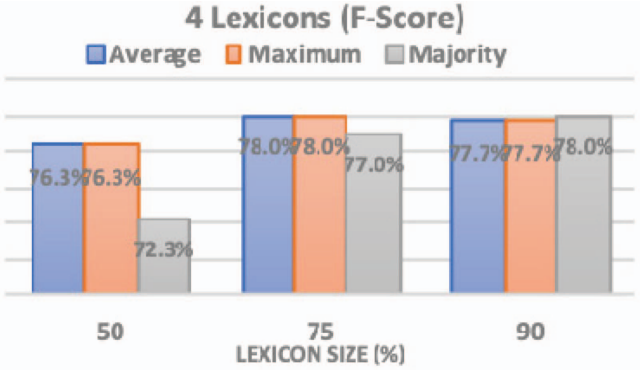

Figure 4 The 4 subset lexicons and three voting methods.

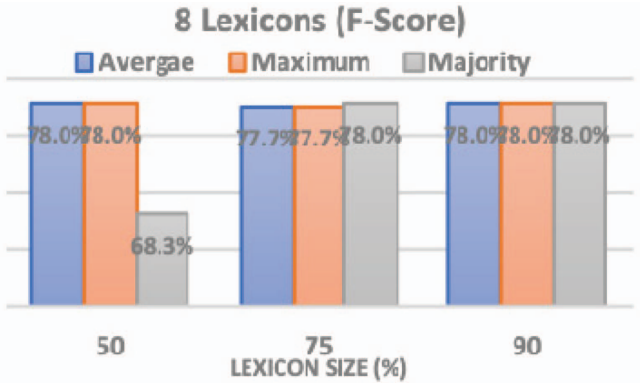

Figure 5 The 8 subset lexicons and three voting methods.

Table 1 SATALex test results

	Test Data			
	Acc (%)	Pre (%)	Rec (%)	F-Score(%)
SATALex	87.3	72.6	82.3	75.8
Best SO Ensemble	82.8	77.3	80.0	78.0

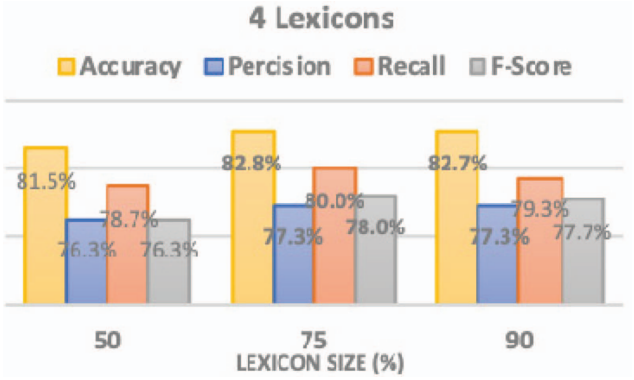

Figure 6 The performance of 4 subset lexicons.

methods; (2) lexicon size 75% using majority voting method; (3) lexicon size 90% using all voting methods. Consequently, we implied from the above graphs that the average voting method in most of the cases produce the highest f-score results. That is why we adopted the average voting method as our decision method.

Table 1 shows the performance measure when using the whole SATALex lexicon for classifying the tweets, while Figures 6 and 7 show the performance measures obtained after running the SO ensembles adopting average voting method using: (1) 4 subset lexicon; (2) 8 subset lexicons.

As shown in the graphs, the best results for all performance measures were produced by the 4 subset lexicons representing 75% of the SATALex lexicon. This subset produced accuracy of 82.8%, precision of 77.3%, recall of 80.0% and f-score of 78.0% with a minor increase compared to other representation percentages for 4 subset lexicons family or when compared to other representation percentages for 8 subset lexicons family. Also, this subset takes less time in building the model and predicting the result as it uses 4 medium sized lexicons. The main benefit of using SATALex lexicon over any other sentiment lexicons is that all the subset lexicons contain domain-related sentiment words with their corresponding adjusted domain polarity.

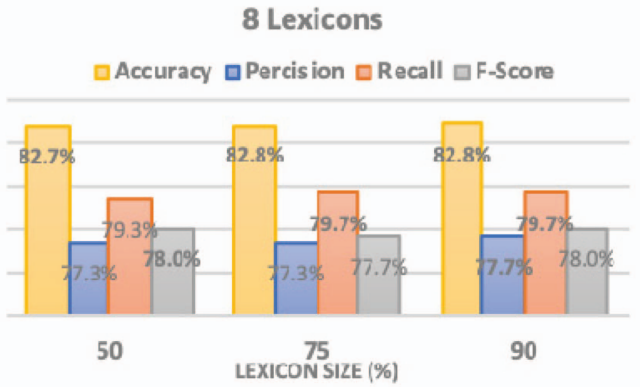

Figure 7 The performance of 8 subset lexicons.

According to the results, the F-score has increased by 2.2% compared to one lexicon SATALex (Shoukry and Rafea, 2019). This increase reflects the fact that there are some sentiment words that cause misclassification of the test tweets implying wrong sentiment. Thus, using subsets from the main lexicon helped in eliminating some of these terms and improving the performance of SO ensemble classifiers. An example of these words is "مشكلة" (problem) which has a negative sentiment in the lexicon. However; these words are mostly used in neutral tweets for general questions. Accordingly, the lexicons which didn't include these terms produced better results. Another example is the term "افضل" which has different meanings depending on the context it is used in. In the modern standard Arabic, it means "best" a positive sentiment term, whereas in the Egyptian dialect it has two different meanings: (1) "remains" a negative sentiment term; (2) "prefer" a neutral sentiment term. On the other hand, creating a subset from the main lexicon by any percentage affects the lexicon's comprehensiveness. Thus, it is possible that some sentiment words are not to be recognized in the test tweets. But the improvements achieved from removing those misleading sentiment words outweighed the loss from not identifying these sentiment words.

Additionally, some tweets hold more than one sentiment. For example:

"Orange_Egypt الجاي عشان الم عامله الوسخه حوله الخطوط البيت فودافون لاورانج الشهر VodafoneEgypt"

meaning:

"Orange_Egypt I will transfer all my Vodafone mobile lines to Orange next month because of bad service VodafoneEgypt"

Table 2 F-Scores for Egyptian test set

	F-Score (%)				
	Set 1	Set 2	Set 3	Set 4	Set 5
SATALex	80.0	72.8	77.1	72.8	**72.7**
Best SO Ensemble	**87.7**	**77.3**	**82.3**	**74.7**	70.3

The sentiment scores for these tweets might sum up to zero as the positive sentiment score cancels out the negative sentiment score, so the tweet gets classified as neutral. However, focusing only the sentiment words of one class will possibly help these tweets to be classified as either positive or negative rather than neutral.

Moreover, we calculated the statistical significance of the proposed SO Ensemble SATALex lexicons (4 lexicons representing 75%). So, we divided the test datasets into 5 sets and calculated the F-Score for each set. The results are shown in Table 2.

Then, we applied the T-Test between SATALex and best SO Ensemble SATALex lexicons using these F-Score values. The value of alpha was set to 0.05. The p-value was 0.018. The difference is significant between SATALex and Ensemble SATALex lexicons since the result is less than the value of alpha.

4.3.2 ML ensemble method

According to the methodology discussed in Section 3.2, we have carried out two experiments to evaluate the performance of the two proposed ML ensemble learning. The first experiment was combining three different classifiers (SVM, NB, ME) on the same training dataset. While the second experiment was combining three SVM classifiers using three different training datasets created by random selection from the main training dataset with replacement. For both experiments, unigrams and sentiment words were used as features to represent tweets. Also, the same test dataset was used in both experiments, together with the decisions taken in Section 3.2.

Table 3 shows the performance of each ML classifier using the two proposed sets of features.

For the results of the ML classifiers, it was observable that significant improvements were obtained using sentiment words for tweets' representation against using unigrams for tweets' representation in all performance measures. These improvements were achieved by using: (1) the ML approach to associate the combination of specific sentiment words to specific class; and (2) the SATALex to identify these sentiment words. SVM and ME had the

Table 3 Test results for each ML classifier

ML	Features	Acc (%)	Pre (%)	Rec (%)	F-Score (%)
SVM	Unigrams	57.0	45.0	35.0	35.0
	Sentiment	**78.45**	**51.0**	**52.0**	**52.0**
NB	Unigrams	56.2	34.0	35.0	32.0
	Sentiment	**64.0**	**54.3**	**41.7**	**35.3**
ME	Unigrams	59.6	42.0	35.0	33.0
	Sentiment	**75.3**	**52.0**	**47.0**	**47.0**

Table 4 Test results for ML ensembles

	Features	Acc (%)	Pre (%)	Rec (%)	F-Score (%)
Ens. 1	Unigrams	60.3	42.0	35.0	32.0
	Sentiment	**71.6**	**55.0**	**50.0**	**48.0**
Ens. 2	Unigrams	59.6	38.0	34.0	31.0
	Sentiment	**78.8**	**51.0**	**52.0**	**52.0**
RFT	Unigrams	57.4	41.0	44.0	41.0
	Sentiment	**80.7**	**69.0**	**80.0**	**72.0**

highest f-score improvements as SVM improved by 17% and ME improved by 14%.

Throughout the experiments, NB showed the least performance than SVM and ME. In fact, the best performance outputs achieved by NB were 64.0% as accuracy and 54.3% as precision. This is because NB is based on probabilities, thus it is more suitable for inputs with high dimensionality. ME achieved the highest accuracy of 75.3%; while its other measures were 1–5% less than the highest values produced by SVM. Concerning SVM, it is considered to be the best performing classifier scoring a significant difference compared to other ML classifiers. In fact, SVM was applied successfully in several sentiment analysis tasks because of its principle advantages. First, they are robust in high dimensional spaces. Second, any feature is relevant. Third, they are robust when there is a sparse set of samples. Finally, most text categorization problems are linearly separable.

Table 4 shows the results of the two proposed ML ensembles using the two proposed sets of features: (1) unigrams; (2) sentiment words. These proposed ensembles were then compared against RFT classifier as suggested by (Shoukry and Rafea, 2019).

Evaluating the results of the two ML ensembles using the two proposed sets of features, it was clear that both ML ensembles had significant improvements using sentiment words for tweets' representation against using unigrams. This result was expected based on the results obtained in Table 3 as

the performance of the ensembles is directly proportional to the performance of the underlying ML classifiers. Consequently, as the performance of the ML classifiers improves, the overall performance of the ML ensemble improves.

Concerning the results of the two ML ensembles, ML ensemble 2 outperformed ML ensemble 1, producing accuracy of 78.8%, recall of 52% and f-score of 52% compared to ML ensemble 1 with accuracy of 71.6%, recall of 50% and f-score of 48%. One possible explanation is that each of the single models may perform well at some parts of the dataset and may overfit at other different parts. Accordingly, each model has different performance on different parts of data. Therefore, by combining these models, the performance of each model tends to improve by reducing the risk of over-fitting. Hence, the performance measures may be improved without affecting the model's predicting performance.

On the other hand, RFT classifier produced better results than the two proposed ensembles using both feature sets. The same observation was presented in (Shoukry and Rafea, 2019) as RFT classifier is proven to be the best according to the literature. RFT classifiers are considered ensemble classifiers for two main reasons. First, they reduce the chances of overfitting by averaging several trees, thus decreasing significantly the risk of overfitting. Second, they cause less variance by using multiple trees, so they reduce the chance of stumbling across a classifier that doesn't perform well because of the relationship between the training and testing data.

Alternatively, if we compared the results of both ML ensembles to the results obtained by SO ensembles using SATALex subset lexicons, it is obvious that SO improves over the ML experiments in all performance measures. For example, considering the best results produced by ML ensembles, the accuracy improved by around 4%. While for the precision, recall and f-score, the improvements were between 26-28%. Based on these outputs, we can confirm that it is recommended to use a SO ensemble for the sentiment classification of Egyptian dialect tweets.

4.3.3 Different domain and SO ensemble

The objective of this experiment is to compare the performance of our proposed SO ensemble method to SATALex lexicon in classifying tweets from 5 different domains (Sports, Politics, Economy, Art, and Communication). Based on the reached conclusion, we have used the SO ensemble method which utilizes 4 subset lexicons with each representing 75% from the original lexicon. Different combining methods were also applied.

Table 5 Test results for SO ensemble

Combining Method	Acc (%)	Pre (%)	Rec (%)	F-Score (%)
Average	**0.71**	**0.56**	**0.49**	**0.50**
Maximum	0.71	0.56	0.48	0.50
Majority	0.71	0.57	0.48	0.50

Table 6 Test results for SATALex lexicon

	Positive	Negative	Neutral	Average
Pre	0.25	0.44	0.74	**0.47**
Rec	0.44	0.40	0.65	**0.50**
F-Score	0.32	0.42	0.69	**0.48**
Acc		**61.63%**		

Table 5 shows the performance of each combination method using the proposed SO ensemble which produced the best result from previous experiment. Table 6 shows the performance of SATALex lexicon in classifying the different domain tweets.

Comparing the obtained results, we can observe that the SO ensemble method improved the f-score measure by almost 2% against using SATALex lexicon. This result reflects the effectiveness of the proposed SO ensemble even when applied on tweets from different domain.

5 Conclusions

This paper has presented a very simple yet powerful ensemble system for sentiment analysis. We began by the SO ensemble which combines multiple SO classifiers using different subsets from SATALex Egyptian lexicon to build more accurate sentiment classifier. Each subset lexicon contributes to the success of the overall system, outperforming single lexicon approach on test dataset. We have also explored the performance of combining three ML classifiers (NB, SVM, ME), and the performance of combining three SVM classifiers using three different training datasets. However, we have concluded that SO ensemble system outperformed the best ML ensemble by almost 26%. The average voting method was used in all ensembles for fair comparison as it proved to be consistently effective and produce the highest results compared to the other combination methods. Comparing our SO ensembles to the presented related works, it was noticeable that our SO ensembles outperform these works by [12–28]%; given that different datasets were used.

This study is part of a bigger project focusing on developing a web application that can "feel" the pulse of the Arabic users with regards to a certain hot topic. This bigger project also includes extracting the most popular Arabic entities from online Arabic content together with the users' comments related to these entities. These extracted popular entities are used to build semantically-structured concepts. It also includes building relations between different concepts and analyzing them to get a sense of the most dominant sentiment. Real case studies are to be conducted to assess the impact of using this tool in increasing performance indicators of industrial and service companies.

For future work, there are different directions for enhancing this work. One direction could be further investigations related to ensemble learning creation and combination methods for better results prediction. One more direction could be exploring and extending the main lexicon (SATALex) using "context embedding". This method focuses on creating more than one-word embedding representation for each word which usually has different meaning in different context. Finally, another direction could be comparing and studying the significance of the proposed methodologies with respect to other languages and dialects.

Acknowledgements

The authors would like to thank ITIDA for sponsoring the project entitled "Sentiment Analysis Tool for Arabic", and the Egyptian industrial company RDI for collecting and annotating tweets.

References

Abbasi, A., Chen, H. and Salem, A., "Sentiment Analysis in Multiple Languages: Feature selection for opinion classification in Web forums," ACM Transactions on Information Systems (TOIS), v. 26, no. 3, pp. 12, 2008.

Augustyniak, Łukasz; Szymański, Piotr; Kajdanowicz, Tomasz; Tuligłowicz, Włodzimierz; Alhajj, Reda; Szymanski, Boleslaw and Kazienko, Przemysław. (2014). "Simpler is better? Lexicon-based ensemble sentiment classification beats supervised methods". In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks

Analysis and Mining (ASONAM 2014), Beijing, China, 17–20 August 2014; pp. 924–929.

Augustyniak, Łukasz; Szymański, Piotr; Kajdanowicz, Tomasz; and Kazienko, Przemysław. (2016). "Fast and accurate - improving lexicon-based sentiment classification with an Ensemble Methods". Conference: 8th Asian Conference on Intelligent Information and Database Systems, 14–16.

Catal, Cagatay; and Nangir, Mehmet. (2017). "A sentiment classification model based on multiple classifiers". Applied Soft Computing, 50 (2017), pp. 135–14.

Deng, Lingjia, and Janyce Wiebe. "Sentiment Propagation via Implicature Constraints." Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, doi:10.3115/v1/e14-1040.

Hamilton, William L., et al. "Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora." Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, doi:10.18653/v1/d16-1057.

Hovy, Dirk. "Demographic Factors Improve Classification Performance." Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, doi:10.3115/v1/p15-1073.

Medhat, Walaa; Hassan Yousef, Ahmed; and Mohamed, Hoda. (2014). "Sentiment Analysis Algorithms and Applications: A Survey". Ain Shams Engineering Journal. 5. 10.1016/j.asej.2014.04.011.

Ohana, Bruno; Tierney, Brendan; and Delany, Sarah Jane. (2011). "Domain independent sentiment classification with many lexicons." In 4th International Symposium on Mining and Web at 25th International Conference on Advanced Information Networking and Applications (AINA), pages 632–637. IEEE Computer Society. doi:10.1109/WAINA.2011.103

Oussous, Ahmed; Lahcen, Ayoub Ait; Belfkih, Samir. (2018). "Improving Sentiment Analysis of Moroccan Tweets Using Ensemble Learning". In: Tabii Y., Lazaar M., Al Achhab M., Enneya N. (eds) Big Data, Cloud and Applications. BDCA 2018. Communications in Computer and Information Science, vol 872. Springer, Cham.

Shoukry, Amira, Rafea, Ahmed. 2012. "Preprocessing Egyptian Dialect Tweets for Sentiment Mining". In *Proceedings of the fourth workshop on Computational Approaches to Arabic Script-Based Languages*. pp. 47–56, San Diego, California, USA.

Shoukry, Amira; and Rafea, Ahmed (2019). "SATALex: Telecom Domain-specific Sentiment Lexicons for Egyptian and Gulf Arabic Dialects". In Proceedings of the 15th International Conference on Web Information Systems and Technologies - Volume 1: WEBIST, 169–176, 2019, Vienna, Austria.

Yang, Yi and Jacob Eisenstein. "Putting Things in Context: Community-specific Embedding Projections for Sentiment Analysis." CoRR abs/1511.06052 (2015).

Zhou, Zhi-Hua. (2012). "Ensemble Methods: Foundations and Algorithms." CRC Press, Boca Raton (2012).

Biographies

Amira Shoukry attended the American University in Cairo (AUC), Egypt where she received her B.Sc. degree in Computer Engineering in 2010. She then obtained her M.Sc. degree in Computer Science in 2013, AUC. Dean's List of Honors, AUC, spring 2012. Her two main publications are "Preprocessing Egyptian Dialect Tweets for Sentiment Mining" and "Sentence-level Arabic Sentiment Analysis". Amira has held different testing and software quality engineering senior positions at IBM Technologies since 2013. She, as a software testing expert and professional, has acquired a solid experience in software quality control of either web, desktop, or mobile applications. She is currently working as a test automation manager at IBM leading some of the major projects. Current Research interests are Data and Knowledge Mining, or Pattern Recognition.

Ahmed Rafea received his PhD from Paul Sabatier University in Toulouse, France. He is a Computer Science Professor and Ex-Chair of the Computer Science and Engineering Department at the American University in Cairo. He served as the Chair of the Computer Science Department and Vice Dean at the Faculty of Computers and Information, Cairo University. He also served as a Visiting Professor at San Diego State University and National University in the United States. Dr. Rafea has led many projects aiming at using Artificial Intelligence and Expert Systems Technologies for the development of the Agriculture sector in Egypt. Dr. Rafea was the principal investigator of several projects for developing Intelligent Systems, Machine Translation, and Social Media Mining in collaboration with European and American Universities. Dr. Rafea's research interests are Data, Text and Web Mining, Natural Language Processing and Machine Translation, Knowledge Engineering and Knowledge Based System Development. Dr. Rafea has authored over 200 scientific papers in International and National Journals, Conference Proceedings and Book chapters.