Emotion detection in Arabic texts extracted from twitter network by using machine learning techniques

Dr.Ammar Alnahhas
Ph.D in Syrian Virtual University
aalnahhas@svuonline.org

Eng. Hiba Mohamed Alkhateeb
Master in Web Science, Syrian Virtual University, Damascus, Syria
hibamoon89@gmail.com

Abstract

Sentiment analysis has recently become increasingly important with a massive increase in online content. It is associated with the analysis of textual data generated by social media that can be easily accessed, obtained, and analyzed. Just a few studies utilized sentiment analysis of social media using a machine learning approach. These studies focused more on sentiment analysis of Twitter tweets in the English language and did not pay more attention to other languages such as Arabic.

The algorithms applied in the previous studies were included :Support Vector Machine Model, Neural Network Model **,**Stochastic Gradient Descent model
4-K nearest neighbors model, Naive Bayes model, Logistic Regression model, Ensemble learning Stacking Method (Random forest, Neural Network and KNN), Ensemble learning Extreme Gradient Boosting Machine (XGBoost) .While the decision tree **(DT)** model was used as a baseline classification comparison with previous best performing models

This study proposes a machine learning model to analyze the Arabic tweets from Twitter. In this model, we apply Word2Vec for word embedding and Several algorithms have been applied in order to reach the highest accuracy in analyzing the sentiments of Twitter users,

The best performing model was obtained in our data set. By getting the best hyperparameters

This is when we use hyperparameter optimization, where we use Grid Search technology

The results show that applying word embedding with an ensemble XGBoost achieved good improvement on average of F1 score and the best of accuracy.

1. Introduction

the current world of technology everyone is expressive in one or other way. People want to express their opinions about various issues be it social, political, economic or business. In this process social media is helping people in a great way. Social networking sites like Facebook, twitter, WhatsApp and many others thus become a common tool for people to express themselves. Analyzing the opinions expressed by the people on different social networking sites to get useful insights from them is called social media analytics. The insights gained can then be used to make important decisions. Among all the networking sites twitter is becoming most powerful wherein people express their opinions in short textual messages called tweets. Analyzing the tweets to retrieve insight information is called twitter sentiment analysis (SA) or opinion mining. Sentiment analysis classifies the sentiment of a tweet into three classes of positive negative and neutral (Pang & Lee, 07 Jul 2008)

Tweets as a social media data source are challenging to analyze. It is written in a slang language that may have grammatical errors which makes it hard for machines to understand (AlZoubi, Tawalbeh, & AL-Smadi, 6 June 2020).

Sentiment analysis has recently become closely associated with the analysis of textual data generated by social media that can be easily accessed, obtained, and analyzed . This encouraged analysts to analyze the public's different opinions on social, health, political, and other issues (Basiri, Abdar, Cifci, & Acharya, 2020).

Challenges in Twitter Sentiment Analysis
The task of sentiment analysis on twitter data is most challenging. The most common challenges associ-
ated with twitter sentiment analysis are as follows:
1. Use of highly unstructured and non-grammatical language in tweets.
2. Use of slang words.

3. Use of sarcasm in tweets
4. Use of words which have subjective context in one sentence and objective in another.
5. Use of negative words to oppose the sentiment of tweet.
6. Use of acronyms and abbreviations.
7. Use of out of vocabulary words.

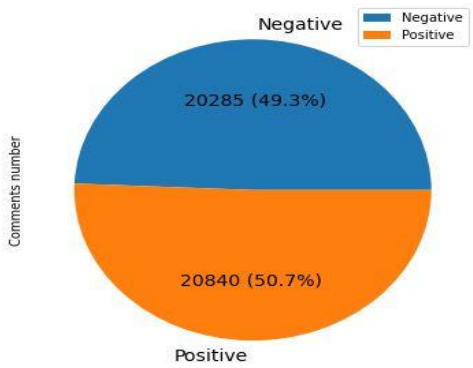
Proposed Arabic Sentiment Analyzer (ASA) Model

The proposed model has several phases, including data collection, preprocessing techniques, feature extraction, and model generation.

2. Data Collection
   Data in the presented Arabic emotions dataset was aggregated from https://www.kaggle.com/.

it was a corpus composed of 41.125 tweets previously collected and labeled for polarity (positive, negative)

The number of positive samples: 20285
The number of negative samples: 20840
as shown in picture



3. Preprocessing Step

The preprocessing is a required step in order to remove the unwanted special and Latin characters. Different techniques were conducted for further analysis as described as follows.

(1)Removing noise from Arabic data: several steps have been conducted in order to remove the noisy complexity of collected data, including removing the punctuations, Arabic diacritics, Numbers, emotions, and no letters from the word such as(*,/ #- / @).

Irrelevant Noise: Tweets, in general, are unclean and contain irrelevant information. (*,/ #- (/). This irrelevant information needs to be cleaned before further processing. i.e:

(2)Normalization: there were several rules applied to normalize the text. For example, each Alif with different forms such as (أ, إ, آ) is replaced with a bare Alif character "ا." Besides, the characters repeated with more than two frequencies and Tatweel "_" were removed.

(3)Tokenization: the tokenization procedure tokened the cleaned text into words in order to filter out unnecessary tokens. The tokenization step is vital before transforming to vectors that are used as input for classifiers.

(4)Streaming: an Arabic light streamer used to remove prefixes and suffixes resulting in reducing the dimensionality of the text data.

(5)Remove stop words: stop words are a set of commonly used words in a natural language occurring frequently and carry less meaning. The natural languages toolkit (NLTK) library has an extensive list of Arabic stop words. The Arabic stop words are integrated and stored in a file. These stop words would take up space in our dataset and valuable processing time; therefore, they were removed.

4. Word Embedding

The function of word embedding is to map closely the words that have a similar meaning or common contexts in the space using word vectors. It efficiently computes word vector representation in the high dimensionality of vector space. The most two common methods used for producing word embedding in NLP are Word2Vec and Global Vectors. The Word2Vec uses a combination of two neural network architectures, including continuous bag-of-words (CBOW) and skip-gram.

Word2Vec effectively computes representations of the word vector in high-dimensional vector space. Word vectors are located in the vector space where terms that have similar semantics and share common contexts are represented in space close to each other. Besides syntactic information, the similarity of word representations extracts semantic features (Pennington, Socher, & Manning, Doha, Qatar, 2014).

So we selected Word2Vec in this study for word embedding.

5.Model Generation

In this section, a brief description of the suitable classification model for enhancing the process of building a model is presented.

5.1. Naïve Bayes (NB)

Naïve Bayes (NB) classifier is a probabilistic classifier that uses the properties of Bayes theorem assuming the strong independence between the features. One of the advantages of this classifier is that it requires a small amount of training data to calculate the parameters for prediction. Instead of calculating the complete covariance matrix, only the variance of the feature is computed due to feature independence. This classifier is widely used as a baseline classifier (Vinodhini & Chandrasekaran, 2012) .

5.2. Support Vector Machine (SVM)

SVM is a machine learning classification technique that uses a function called kernel to map a space of data points in which the data is not linearly separable onto a new space, with allowances for erroneous classification. It has been successfully utilized on Arabic sentiments analysis (Abdulla, Al-Ayyoub, & Al-Kabi, 2014.) .

5.3. Logistic Regression (LR)

The logistic regression classifier (LR) calculates the conditional probability distribution of a class relied on the dataset. It supposes no former knowledge. The ratio of the training dataset puts a constraint on the conditional distribution, thus forcing the classifier to find the ME distribution that is consistent with the constraint (S. Dreiseitl & L. -Machado, 2002). Since the classification is binary and our dataset is small in size, LRCV is utilized with a solver (liblinear) to discover the weights of the parameter to minimize a cost function. In this paper, the experiments are conducted using 10-fold cross-validation.

5.4. Stochastic Gradient Descent (SGD)

Stochastic gradient descent (SGD) is an optimization technique for converging on a problem solution by choosing an arbitrary solution. It measures the goodness of fit under a loss function and iteratively takes steps in the direction that minimizes loss. According to (A. Ghallab, A. Mohsen, & Y. Ali, 2020), this classifier has some advantages such as efficiency and ease of implementation. Thus, it is applied in this paper.

5.5.Ensemble approaches

Ensemble approaches combine two or more from the available approaches. This is in order to come up with enhanced results. Ensemble approaches provide greater generalization and enhanced performance compared to stand-alone approaches. Ensembles can get benefits from the combination of all approaches to enhance accuracy (OmarAlZoubi, Tawalbeh, & AL-Smadi, 16 October 2020).

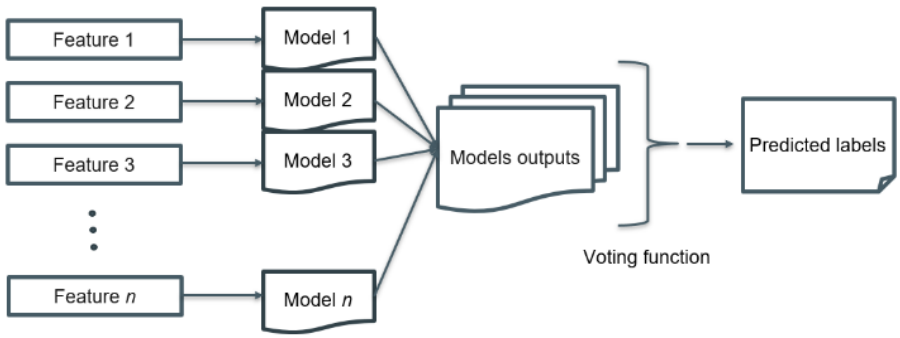


Figure 2: The scheme of the ensemble architecture

5.6. Random Forest (RF)

Random forest is a type of supervised machine learning algorithm based on ensemble learning. It combines multiple decision trees and then produces a forest of trees, hence the name “random forest.” The random forest algorithm can be used for both regression and classification tasks.

5.7.Voting

A voting classifier is an ensemble classification method that has the advantage of combining the predictions by majority voting from multiple machine learning algorithms. It exploits the features of each algorithm (M. Swamynathan, 2019). Several combinations of single classification methods are used with the majority vote in this paper. The voting classifier combined RF, SGD, SVM, BNB, and LR.

5.8.XGboost

XGboost (XGB) Chen and Guestrin (2016) is a popular method used to solve data science problems. XGB is considered as a distributed gradient boosting library prepared to be highly efficient. It is a flexible library dedicated to working with NLP problems such as classification, regression, and ranking problems. There are several common reasons to use XGB. First is Regularization; Gradient Boosting Models do not include regularization implementation. However, XGB is a regularized boosting technique used to reduce over-fitting. Second is Built-in Cross-Validation; it allows running the cross-validation once for each iteration of the boosting process to get accurate optimum number of boosting iterations. Third is Performance; it dominates structured classification and regression problems. Fourth is Speed; it is fast compared to other gradient boosting. Finally, Handling Missing Values; XGB is different from neural networks where it has a built-in routine used to handle missing values. XGB Regressor was used to obtain the intensity of a given tweet

(emotion) (OmarAlZoubi, Tawalbeh, & AL-Smadi, 16 October 2020).

6.Results and discussion

Table 6 presents the results of each model computed for each emotion and their macro-average results.

	Model	Test Accuracy	Recall	Precision	F1 score
1	Random forest	0.79155	0.73826	0.83481	0.78357
2	Support Vector Machine	0.79689	0.73293	0.84904	0.78672
3	Neural Network Model	0.77516	0.75946	0.79208	0.77542
4	Stochastic Gradient Descent	0.66373	0.61915	0.69085	0.65304
5	K nearest neighbors	0.80142	0.73237	0.85833	0.79036
6	Naive Bayes	0.65335	0.60739	0.68017	0.64172
7	Logistic Regression	0.69323	0.664	0.71537	0.68873
8	Ensemble learning Stacking Method (Random forest, Neural Network and KNN)	0.80801	0.76867	0.84194	0.80364
9	Ensemble learning Stacking Method (Neural Network and KNN)	0.80664	0.76723	0.84056	0.80222
10	Ensemble learning XGBoost	0.85146	0.83904	0.8658	0.85224

. The XGB outperformed the other all techniques and achieved **F1 score**=0.85224

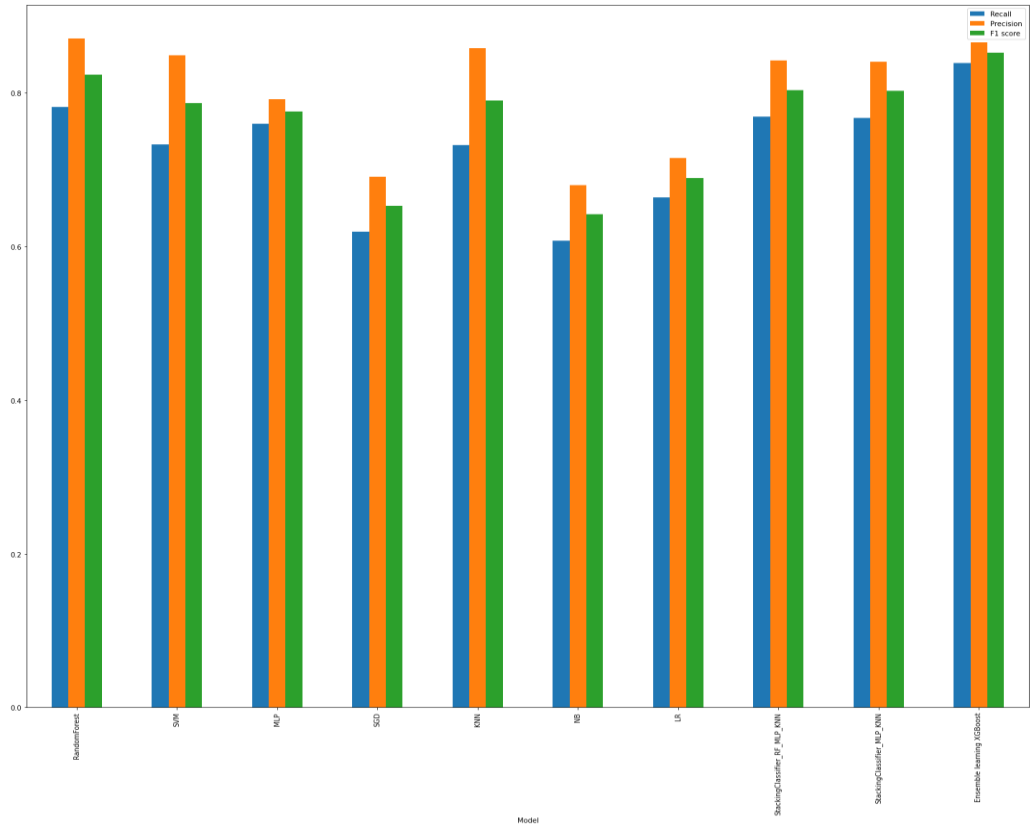
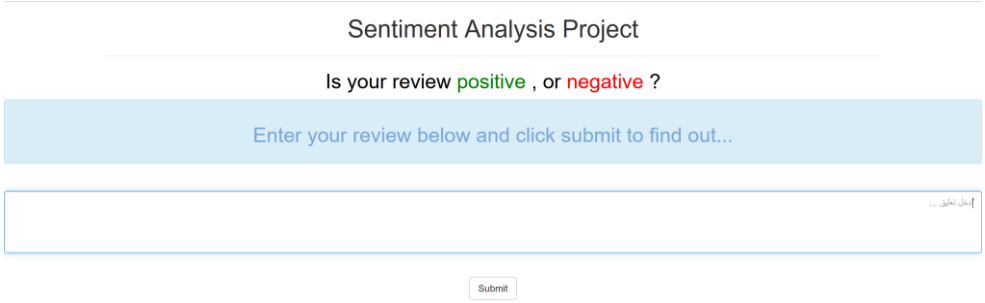


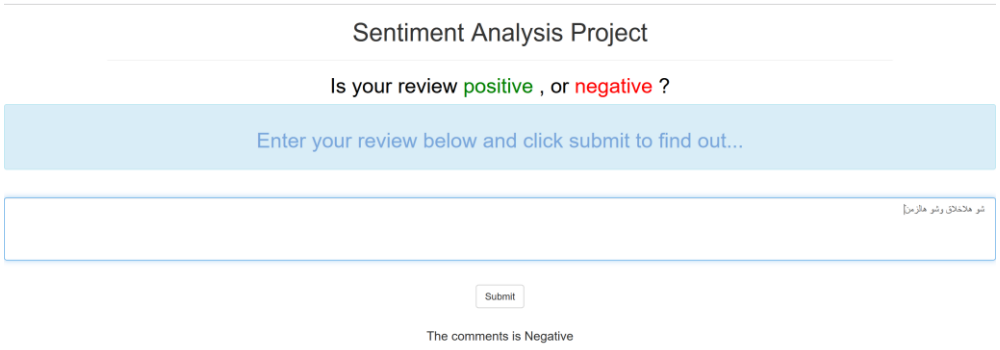
Figure 3

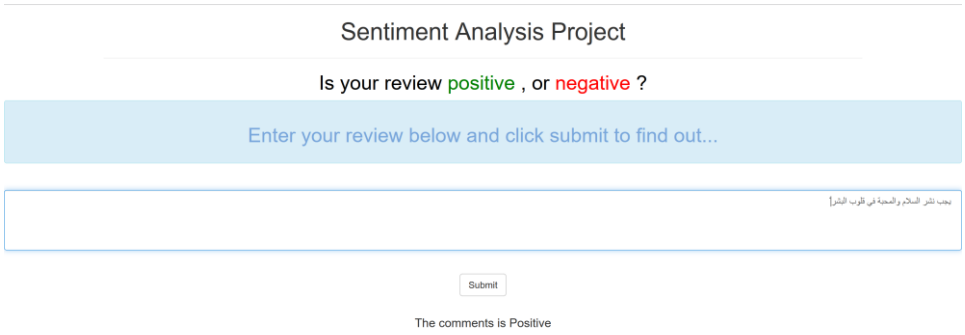
A website has been created in the following languages (Html.Css.Java Script) through which the user can write the comment and show him the result of the comment, positive or negative:

figure 3 shows the average *F*1 score te**st Accuracy** for all classifiers.



We notice that when the user entered a comment, whether negative or positive, the system analyzed it and gave it the result as having negative content or positive content.
We note in the following two pictures the application on it.





7.Conclusion

In this paper, we introduced a Twitter based dataset for Arabic emotion detection,

periments with this dataset have shown that the Complement XGBoost classifier yields the best results with an overall accuracy of 85.1%.

how promising results, more efforts are needed to achieve better results. As part of our future work, we intend to experiment with deep learning ap-

proaches, which have been very successful in English sentiment analysis. We also

plan to expand our dataset to include more diverse data from multiple dialects

8.Referenc

Abdulla, N. A., Al-Ayyoub, M., & Al-Kabi, M. N. (2014.). An extended analytical study of Arabic sentiments. *International Journal of Big Data Intelligence*, 103–113.

Pang, B., & Lee, L. (07 Jul 2008). Opinion Mining and Sentiment Analysis. *now the essense of knowledge*, 1-135.

Pennington, J., Socher, R., & Manning, C. D. (Doha, Qatar, 2014). Glove: global vectors for word representation,. *Empirical Methods in Natural Language Processing (EMNLP),*.

Vinodhini, G., & Chandrasekaran, R. (2012). Sentiment analysis and opinion mining: a survey . *International Journal,*, pp. 282–292.

A. Ghallab, A. Mohsen, & Y. Ali. (2020). Arabic sentiment analysis: a systematic literature review. *Applied Computational Intelligence and Soft Computing*, 21 pages.

AlZoubi, O., Tawalbeh, S. K., & AL-Smadi, M. (6 June 2020). Affect detection from arabic tweets using ensemble and deep learning techniques. *Journal of King Saud University - Computer and Information Sciences*.

Basiri, M. E., Abdar, M., Cifci, M. A., & Acharya, U. R. (2020). A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques,. *Knowledge-Based Systems,*, 198-200.

M. Swamynathan. (2019). Mastering Machine Learning with Python in Six Steps:. *A Practica Implementation Guide to Predictive Data Analytics Using Python*.

OmarAlZoubi, Tawalbeh, S. K., & AL-Smadi, M. (16 October 2020). Affect detection from arabic tweets using

ensemble and deep learning techniques. *Journal of King Saud University - Computer and Information Sciences*.

S. Dreiseitl , & L. -Machado, O. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, 352–359.

Pennington, J., Socher, R., & Manning, C. D. (Doha, Qatar, 2014). Glove: global vectors for word representation,. *Empirical Methods in Natural Language Processing (EMNLP),*.

Vinodhini, G., & Chandrasekaran, R. (2012). Sentiment analysis and opinion mining: a survey . *International Journal,*, pp. 282–292.

AlZoubi, O., Tawalbeh, S. K., & AL-Smadi, M. (6 June 2020). Affect detection from arabic tweets using ensemble and deep learning techniques. *Journal of King Saud University - Computer and Information Sciences*.

Basiri, M. E., Abdar, M., Cifci, M. A., & Acharya, U. R. (2020). A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques,. *Knowledge-Based Systems*,, 198-200.