Emotional analysis & explainable AI in Arabic dataset

Alaa Mahmoud Mohamed Mahmoud
Student at faculty of computing and data science
cds.AlaaMahmoud67249@alexu.edu.eg
Karim Anani Atteya Ismail
Student at Computer and Systems Department
Faculty of Engineering
Karim.atteya2023@alexu.edu.eg
Mirna Maher Mahmoud Abdeaziz
Student at faculty of computing and data science
cds.MirnaMaher81685@alexu.edu.eg
Mahmoud Alaa Mahmoud Abdelfattah
Student at faculty of computing and data science
cds.mahmoudalaa15074@alexu.edu.eg

Abstract → ***With rapid development of social media, there is large variety in feelings, thoughts and emotions about everything in which it varies from person to person***
Emotion Analysis and Sentiment Analysis both help to revolutionize the way people respond to new products
Developments in Arabic information retrieval did not follow a high improvement
Studies on topic models, which provide a good way to automatically deal with texts, are not complete enough to assess the effectiveness of the approach on Arabic texts
The purpose of this study is to use explainable AI to understand performance on datasets using different ML models on Arabic datasets.
A combination of Arabic text preprocessing techniques were tested with word embedding, machine learning and LIME tool for explainability
The ensemble model between (LR,SVM,CNB) get the best accuracy of 67% compared to other ML classifiers

I. INTRODUCTION

Emotion in social media has been the subject of considerable research and media although researchers have made numerous efforts to use the emotions we express in status updates to make inferences about our emotional lives an attention, Emotion analysis is the process of identifying and interpreting the underlying emotions expressed in textual data.
Emotion analytics uses text data from a variety of sources to analyze it and understand the emotions that underpin it,
For example: They help you to understand your audience, social data can help you create better content, help you understand competitors.
Sentiment analysis (SA) is the process of detecting positive or negative sentiment in text. It's often used by businesses to detect sentiment in social data, gauge brand reputation and understand customers, since humans express their thoughts and feelings more openly than ever before, sentiment analysis is fast becoming an essential tool to monitor and understand sentiment in all types of data.
Emotion classification, or emotion categorization, is the task of recognizing emotions.
The problem is when Arabic texts include many translated and transliterated named entities whose spelling in general tends to be inconsistent in Arabic texts same word may have many different meaning due to the Diacritics , many researchers have improved a lot in English text, so the Arabic text needs to be improvements

The purpose of this study is to use explainable AI to understand performance on datasets using different ML models on Arabic datasets.
Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to trust the results and outputs created by machine learning algorithms. Many people have distrust in AI, yet to work with it efficiently, they need to learn to trust it, that's why we need to use XAI.
We use different combination of preprocessing technique, word embedding model, classical machine learning models and using LIME tool for explainability.
Our contribution is to use the previous work, then use our tool (LIME) to explain and understand dataset and get the words that effect on it to classify them into the corresponding category. Fear, Happiness, Sadness, Love, joy, angry, sympathy and none are all examples of emotions that can be detected and classified in a text.

II. RELATED WORK

The following is a brief review to the emotion classification of Arabic Twitter data related to previous work.
The size of dataset from the Nile University was 10,065 and has been annotated manually with eight emotions: sadness, anger, joy, fear, surprise, love, sympathy and none. The term "none" was used to label neutral tweets. The tweets were collected using the" Olympics" hashtags from Egypt in the period between Jul 2016 and Aug 2016.
Ahmed El-Sayed [8] used support vector classifier, Logistic regression, Complement Naive Bayes classifier, Simple Naïve Bayes classifier, AdaBoost, Decision Tree, Sequential Minimal Optimization, Random Forest and Weighted Voting Ensemble method, the best results were obtained using a Weighted Ensemble method with an overall accuracy of 69.7%
A comparison of the related work is presented in Table I, II and III, the related work revealed that there aren't any explanations about how the model works, so we decided to extend the work by explaining the results of machine learning algorithms using Explainable AI methods, comparing results of different preprocessing techniques using Lime and making some conclusions about our models, extracting the words which have high weights that affect our machine learning algorithm on each emotion.

TABLE I: The pre-processing techniques used in the literature

Pre-processing	2021	Ours 2022
Normalization	√	√
Removal of stop words	√	√
Removal of non-Arabic letters and spaces	√	√
Diacritics removal	√	√
Links, mentions, and retweet indicators removal	√	√
Remove suffixes and prefixes	√	x
Disapprobation words	√	x
Reducing words to their roots	√	√

Occurrence removal	√	√
Negation patterns	√	X
Punctuation	√	√
Replace emoji's and emoticons with emotions	√	√

III. APPROACH

Given the limitations of the reviewed literature in using machine learning models and modern word embedding for Arabic Emotion classifications, we aimed at studying and analyzing the behavior of different classifiers with different preprocessing techniques to have a better understanding of the relation between a word and a prediction probability.

TABLE II: Emotion analysis related work summarized

Features			2021	ours 2022
Dataset	Tweets		√	√
Feature extraction	TF-IDF		√	√
Emotion Classification	Classical Machine Learning	SVM	√	√
		Naive Bayes	√	√
		Complement Naive Bayes	√	√
		Logistic Regression	√	√
		AdaBoost	√	X
		Decision Tree	√	√
		Random Forest	√	X
		Sequential Minimal Optimization	√	X
		Ensemble Voting	√	√
Performance Evaluation	Accuracy		69.7 %	67%
	Precision		69.8 %	67.3%
	Recall		69.5 %	65.6%
	F-score		69.2 %	65.1%

For this, multiple combinations of text preprocessing techniques, shown in Table (I), state-of-the-art feature extraction and word embedding models such as TF-IDF were tested and ensemble classifiers were considered, and Local interpretation model (LIME) was chosen as our explainable AI method.
These new experiments were conducted on the dataset provided by the Nile University [7] The details of the methods used are described in the following subsections.

A. Dataset

The dataset was collected by a research group at Nile University (NU) [7]. It is a balanced dataset consists of 10,065 Arabic tweets mostly using Egyptian dialect, and was manually annotated using eight emotions (sadness, anger, joy, surprise, love, sympathy, fear and none).

It was collected using The" Olympics" hashtags. The dataset was split into 70% for Training and 30% for testing.

B. Experimental design

Data preprocessing is important to minimize the noise and get a better classification accuracy.
The experiment was started by using the preprocessing techniques in Table [I] and testing the impact of adding a new one to replace emojis with its equivalent Arabic meaning.
Emojis were grouped in a dictionary and were replaced with the equivalent Arabic meaning then unneeded words were removed.
In terms of feature extraction, Term Frequency-Inverse Document Frequency (TF-IDF) model was implemented and was used along with the proposed preprocessing techniques.
Classical machine learning techniques were tested namely: Naive Bayes, Complement Naive Bayes, SVM, Logistic Regression and Decision Tree as illustrated in Table [II.]
Finally, a weighted voting ensemble was made between classical machine learning models (Logistic Regression - SVM - Complement Naïve Bayes)
It was built by assigning greater weights to the best classifiers.
To evaluate the performance of different classifiers, accuracy, precision, recall and F-score were used.
For each classifier, the error metrics were calculated for each individual label (joy, sadness, sympathy, anger, fear, surprise, love, none).
The final error metrics were calculated as the average of all the metrics for all the labels.
Regarding the explainable AI part, local interpretable model-agnostic explanations (LIME) was used to compare the behavior of different classifiers and analyze the relation between words used in tweets and significant emotions.
Studying the behavior of classifiers and analyzing its features using explainable AI would be beneficial in order to ensure that the model prediction is reliable.

IV. RESULTS AND DISCUSIONS

We have made 6 main experiments on our project
(All preprocessing steps have been fixed and we just start to deal with 2 steps only which are stemming and dealing with emojis.)

Remove emoji without stemming [1]:

We start with removing emoji from all rows and try to not use stemming which returns each word to its root, to follow up the results of lime and see their effects on the visualization.
We used an accuracy measure to compare results and found the result of these steps is good with our ML models as a beginning and with lime.
The accuracy of SVM is 65 %
The accuracy of the ensemble model is 64%

(We try other experiments to improve these results)

Remove emoji with stemming [2]:

We tried removing the emoji but with stemming to conclude the effect of stemming on results and we found that accuracy increases up to 3 % in each model we use and it is a good step.

The accuracy of SVM is 67%

The accuracy of ensemble model is 67%

With emoji, with stemming [3]:

We start to deal with emoji and leave as it is in data with stemming for text.

The accuracy of SVM is 65 %

The accuracy of the ensemble model is 66 %

So we conclude that models don't correctly deal with emojis, so we will try to replace them with a word that reflects their meaning.

With emoji, without stemming [4]:

Try to find if stemming has a high effect on the accuracy or just emojis,we try to leave emojis without doing any type of stemming on it

Therefore, stemming is a good preprocessing step as it increases the accuracy of experiment [2] when we use it and decreases it in experiment [4] when we do not use it.

The accuracy of SVM is 64 %

The accuracy of the ensemble model is 64 %

Replace emoji without stemming [5]:

As we say before we will replace each emoji in the dataset with its meaningful text but without stemming

We found the result of these steps is good with our ML models and with lime but not the best one so we try the last experiment.

The accuracy of SVM is 66%

The accuracy of ensemble model is 65%

Replace emoji with stemming [6]:

The accuracy of SVM is 67%

The accuracy of the ensemble model is also 67%

(We found the best results in this experiment so we built the remainder of the project on it.)

All these experiments were done on lime in Table [Ⅴ] and we notice that when we replace the emoji with text the probability of the words affecting the choice of the class is increased and the explanation is better

Discussion

Comparing the common models with the previous work on the same dataset:

SVM 68, 67 %, Naive Bayes 62, 60%, Complement Naive Bayes 64, 63%, Logistic Regression 69, 66%, Decision Tree 57, 56%, Ensemble 69, 67%

V. CONCLUSION AND FUTURE WORK

The objective of the paper is to explain the results of machine learning models using Lime.

We were able to detect the most words that affect each annotation using weighted average probability of correctly classified data in the test set.

In the future, we consider using different Explainable AI methods to gain more insights about how our models take decisions, considering the context of emoji in the tweets rather than replacing it, applying pre-trained models

For example: Arabert and Marbert, making framework that can detect mistakes in the manual annotation process and correct it and experimenting the same methods with misclassified data in the test set.

TABLE III: Emotion analysis related work results

Classifiers	TF-IDF	
	2021	2022
Logistic regression (LR)	68 %	66 %
Support vector machine (SVM)	68 %	67 %
Multinomial naïve Bayes	61 %	60 %
Complement naïve Bayes	64 %	63 %
Decision tree (DT)	54 %	56 %
ENSEMBLE (LR,SVM, CNB)	69 %	67 %

TABLE Ⅳ: RESULTS WITH EXPERIMENTS

CLASSIFIERS	**emoji, stem**	**No emoji , stem**	**emoji ,no stem**	**No emoji ,no stem**	**Replace emoji , stem**	**Replace emoji ,no stem**
Logistic regression (LR)	65 %	66 %	64 %	64 %	66 %	64 %
Support vector machine (SVM)	65 %	67 %	64 %	65 %	67 %	66 %
Multinomial naïve Bayes	60 %	60 %	59 %	59 %	60 %	59 %
Complement naïve Bayes	62 %	63 %	62 %	62 %	63 %	62 %
Decision tree (DT)	56 %	55 %	53 %	52 %	56 %	54 %
ENSEMBLE (LR,SVM, CNB)	66 %	67 %	64 %	64 %	67 %	65 %

TABLE V: LIME results for one tweet

	emoji with stem	no emoji, stem	emoji ,no stem	no emoji ,no stem	Replace emoji ,stem	Replace emoji ,no stem
Complement NB	27 %	23 %	30 %	26 %	26 %	33 %
Multinomial NB	27 %	49 %	30 %	41 %	63 %	64 %
SVM	100 %	99 %	100 %	99 %	99 %	100 %
LR	92 %	91 %	85 %	76 %	95 %	92 %
Ensemble(LG,NB, SVM)	97 %	96 %	95 %	93 %	97 %	97 %

REFERENCES

[1] https://colab.research.google.com/drive/1OH557_UaDdZk5jmXnV9FVCUDYxVxmVRK

[2] https://colab.research.google.com/drive/1N9QOzJAZvu3FjRuAnYFOyhcJN4Kfq3gX

[3] https://colab.research.google.com/drive/1YQhbE0DXqpTVs-crEZqQsH6Mld9i3aIY

[4] https://colab.research.google.com/drive/1-rLtJK163YLjliFR9jx2QIZfZtQApf5a

[5] https://colab.research.google.com/drive/1ZwUGEFC-fj3YoOCA80dd6_oqdEM9m9q6

[6] https://colab.research.google.com/drive/1vlfj5CL6VNF3fD8sU2lwvWrqfp7aZ4Kr

[7] https://docs.google.com/spreadsheets/d/1i0RC0p4omS3grwXaeq7TgKlkCCdOEyOk5GoStQ1dlqw/edit?usp=sharing

[8]https://drive.google.com/file/d/1AXnAuJM06fFz_qeRADXIBFxjQuNBTryj/view?usp=sharing