Available online at https://www.asjp.cerist.dz/en/PresentationRevue/134

RIST

Information Processing at the Digital Age Journal

CERIST Natural Language Processing Challenge

March 29th, 2023

Arabic Sentiment Analysis within COVID-19

Slimane Arbaoui, Alaa Eddine Belfedhal

Higher School in Computer Science 08 May 1945, Sidi Bel Abbes, Algeria.

Abstract

In this paper, we give a brief study that allow us to analyze some Arabic tweets posted in the Covid-19 period and classify them into "Positive, Negative and Neutral". This paper is about our participation on CERIST Natural Language Processing Challenge. We worked on a dataset that consist of 4800 tuples on which we applied three different approaches "Naive Bayes, Neuron network and Stochastic gradient descent (SGD)" where the last algorithm gave the best result with an accuracy of 91%.

Keywords: Covid-19, Arabic sentiment, Classification, Text analysis.

1. Introduction

In the Covid-19 period, people spent almost all their time in front of PCs and smart phones and started to share their feelings and sentiments about this pandemic. At that time, researchers started to see that their posts and tweets collected and stored in datasets that can be used to train models which predict emotions and classify them. Most of the models and algorithms proposed by these researchers are English based and for that we tried to build a model that can classify Arabic tweets into labels "Positive, Negative and Neutral".

This project came under the opinion mining and sentiment analysis task proposed within the CERIST Natural Language Processing Challenge. This challenge provides us datasets that we can use to train a machine learning model also it gives us a test dataset to measure the performance of it , Hadj Ameur and Aliane, 2021. In the next sections, we will introduce some of the related work, the dataset produced by CERIST and how we manage to balance it. Also, we will define how we manage to organize the data before

fitting it into a machine learning model. We will define the models used in this study and compare them based on the accuracy.

2. Related work

In this section we will provide an overview of researcher work in this topic. Starting with the work of Folorunsho, 2020, where the author used a dataset consisting of 1800 tweets labelled as positive and negative to perform a classification task where he used four different models "Naïve Bayes, Logistic Regression, Random Forest and Support Vector Machine" that gives promising results which exceed 84%. He also mention that the difference between Arabic and English NLP is the pre-processing steps where he demonstrate some of them.

In Ali, 2021,the author used two datasets collected during the Covid-19 pandemic, and where he used an effective preprocessing technique and various Machine Learning algorithms and compare them. The best accuracy was achieved by SVM-based models with more than 89.1%.

El-Masri et al., 2017 presented a web based tool that can classify Arabic tweets into (negative, positive, both, and neutral) and offer to the user the possibility to choose the parameters that the classification will based on, like: time of the tweets, preprocessing methods, n-grams features and Lexico-based methods. This tool used a model that was trained with 8000 random samples where they found that Naive Bayes are good in predicting topic polarity with an accuracy of 70%.

3. Overview of Dataset

In this section, we get to know the dataset that we will work with. This dataset as we mention before has been built by CERIST from social networks and also data from other datasets, Hadj Ameur and Aliane, 2021. It focuses on Covid-19 related posts. CERIST provide us with two tsv files:

- Train_aracovid_sentiments.tsv: used to train our model
- Test_aracovid_sentiments.tsv: used to test our model

We used Pandas to visualize the first tuples of our dataset and we found the result below (Figure 1). We see that we have two main columns “Text” and ”sentiment” where the first one is our input and the second one is the result or the class, we also find that tuples are ordered bases on the class so we introduce some random order in it. We see the different values that can be used in the sentiment column [“Positive”, ”Negative”, “Neutral”], these values have the following distribution (Figure 2, Table 1)

Table 1. Data Distribution

Value	Tuples
Neutral	1600
Positive	1571
Negative	957

index	Text	sentiment
0	علاج فايروس الكورونا" كل يوم الصبح الواحد يقرط ثلاث حبات ثوم، ومع انو الثوم ما إله علاقة بفايروس الكورونا بس هكذا بتضمن لا حدا يقرب عليك ولا حدا يبوسك . 😂😂😂 😂😂😂 الله يبعد عنا وعنكم البلاء يارب"	Positive
1	. هههه هذا مكان اخطر من كورونا فايروس 😂😂🔥.	Positive
2	💔□😐 خليه يجيبوا لقاح الكورونا وما يكفينا ويقوموا يزيدوه مي ويخضوا الطبه. الاردن ويعرفها	Positive
3	الطواقي وتجار الطرد لحد يطلعون اشاعات ههههههه قمه الغباء #اخ_حالات_كورونا_تكاهر_القصر	Positive
4	@Waleed_Albashaa ههههههه انتظر بعد ما تخلص كورونا عادي 😂	Positive
5	♥♥ منع_التجول يارب نفس ما جاء هذا الوباء فجأة بقدرة الله وغير نمط حياتنا يروح فجأة وبقدرة الله وحده يترجع حياتنا نفس أول وأحسن يارب نرجع نحس بالصلاة ونرجع لجمعة الاهل اللي مليانين شوق ليهم#	Positive
6	😂😂قالو جاي اعصار وفيضان يعني اللي ماجاه كورونا واللي ماطلش اسمه فالسحورين اينتيه السيل	Positive
7	ههههههه الكاضمي وتراكمي اخو فكرة عن اللي يريدون تراكمي وشوفوا إملاكيم شوقوا الاخطاء والسنين شوفوا منو يريد تراكمي 😂😂🌚 التراكمي مرفوض بجماعة مرفوض كافي تلحون وخلينا نطالب @DRBROFE بالتأجيل حفاظاً على ارواحنا وارواح أهلنا #ابادة_جيل_تنتهي_بالتأجيل #التأجيل_مطلبنا_كورونا_لن_ترحمنا https://t.co/vWA71E8u59	Positive
8	- عقلين ايه ف بلكم اللي ملياة كورونا دي ؟ = جوازات .. خطوبات .. ارتباطات .. كتيييير اوي - وش خايفين تتعدوا ؟ = هههه !!!! ماتسمعونا زغروطة #فيروس_كورونا - https://t.co/TnIVbC1WlH	Positive
9	♥ هذا ما عرفة بيابي #كورونا بلاء من الله وَلَنَبْلُوَنَّكُم بِشَيْءٍ مِّنَ الْخَوْفِ وَالْجُوعِ وَنَقْصٍ مِّنَ الْأَمْوَالِ وَالْأَنفُسِ وَالثَّمَرَاتِ ۗ وَبَشِّرِ الصَّابِرِينَ (155) اللهم نسألك الصبر و العفو والعافيه .. 🙏 #صباح_الخير https://t.co/HjCssKO9g2	Positive
10	@hamedalhinai02 هههه بعد كورونا إن شاء الله اما حالياً ملتزمين بالبيت 😂👍	Positive
11	كورونا سيذهب الوباء ومعه البلاء بإذن الله تفاءلوا بالخير تجدوه#	Positive
12	هل تعلم لماذا سميت سورة يوسف بـ (أحسن القصص) ؟ لأنها تعلمنا ان : السجين سيخرج ، و المريض سيشفى ، والغائب سيعود ، والحزين سيفرح ، والكرب سيزول ، و ان ابتلاء المؤمن كله خير ، فـ مع الله لا يخيب رجاء ، فلا تيأس ، و ثق بالله . #كوفيد_19	Positive
13	@dabbas80 ههههههههههه لحد أخطر من كورونا عالكوكب 😂😂	Positive
14	الحراك_مستمر# #hirak #algérie قال لأمه الطاعنة في السن " راهي دخلت كورونا للجزائر " ردت عليه :" ضرك يدوها غير أصحاب المرسيدة ولكلاب حنا ما يمسحنا والو مع ولاد لحرام"😂 #فيروس_كورونا #الجزائر	Positive
15	قريباً .. إن شاء الله ستختفي الكمامة وتشرق الإبتسامة كما وأثق بالله . #كورونا #كورونا_السعوديه	Positive
16	واعلم أن ما أخطأك لم يكن ليصيبك، وما أصابك لم يكن ليخطئك، واعلم أن النصر مع الصبر وأن الفرج مع الكرب وأن مع العسر يسرا #كورونا #السعودية #كوفيد_19 #مساء_الخير #الأردن (إِنَّمَا أَشْكُو بَثِّي وَحُزْنِي إِلَى اللَّهِ) نعود_بحذر #كلنا_مسؤول# #islamic #quran #trending #الامارات https://t.co/LapgFBg3kN	Positive
17	@xxx_6445 الجلب قبل كورونا ليست الجلب بعد كورونا هههه	Positive

Fig. 1. The first look of Arabic Sentiment Dataset

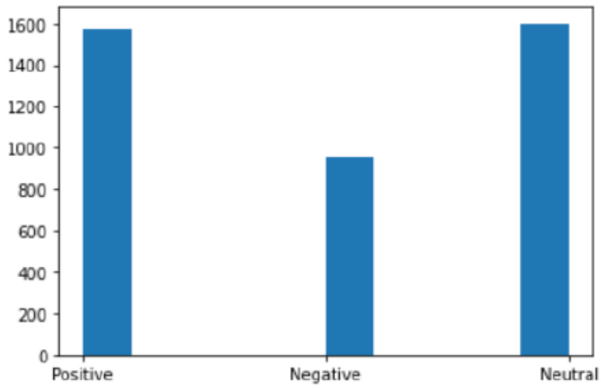

Fig. 2. Data Distribution

We add some negative tuples to balance our data and avoid problems in the future, to do that we use the over-sampling technique that gives the result shown in Table 2

Table 2. Data Distribution after the over-sampling

Value	Tuples
Neutral	1600
Positive	1600
Negative	1600

At the end we finish with a dataset consisting of 4800 tuples and they are well distributed.

4. Data preprocessing

As shown in the previous section, the data contains some non-Arabic characters like Emojis and URLs that need to be cleaned. Hence, we introduce a function (Figure 3) that returns a clean text following the steps below:

- Deleting numbers and extra spacing.
- Extraction of words.
- Deleting stop and long words where we define a long word as a word with more than fifteen characters based on what we find in Google about the longer Arabic word*.
- We used also the Qalsadi† library to extract to lemmatize words and delete some stop words that we did not mention in our stop words documents.

```
def pre1(x):
 tokenizer = RegexpTokenizer(r'\w+')
 f = tokenizer.tokenize(x)
 resultwords = [word for word in f if word not in stwords and len(word)<=15]
 if(len(resultwords)!=0):
    result = ' '.join(resultwords)
    text_nonum = re.sub(r'\d+', '', result)
    text_no_doublespace = re.sub('\s+', ' ', text_nonum).strip()
     lemmas = lemmer.lemmatize_text(text_no_doublespace , return_pos=True)
    words = []
    for j in lemmas:
        if (type(j) is tuple):
        if (j[1] != "stopword" and j[1] != "all"):
        words.append(j[0])
    result=" ".join(words)
    return result
 else:return ""
```

Fig. 3. Preprocessing Function

* Arabic Longest word: https://www.uaemoments.com/what-is-the-longest-arabic-word-331501.html
† Qalsadi Library: https://pypi.org/project/qalsadi/

5. The model

In this section we define the pre-existing algorithms that we used to build our model starting with

5.1. Naive Bayes

Naive Bayes is a simple probabilistic classifier and Bayesian network model that assumes the independence between the features. This kind of classifier is highly scalable and require number of parameters equal to the number of features which are in our case the number of words. According to (Loukas, 2020) "Naive Bayes classifiers have been heavily used for text classification and text analysis machine learning problems", and that why we use it.

This classifier gives a promising result with an accuracy of 90.5% as shown in the next table (Table 3).

Table 3. Naive Bayes result

	Precision	Recall	F1-Score	Support
Negative	0.95	0.92	0.93	652
Neutral	0.85	0.93	0.89	631
Positive	0.92	0.87	0.89	637
Accuracy			0.90520	1920

5.2. Artificial neural networks (ANNs)

Computing systems inspired by the biological neural networks that make up animal brains, they are commonly referred to as neural networks (NNs) or neural nets. Artificial neurons, which are a set of interconnected units or nodes that loosely resemble the neurons in a biological brain, are the foundation of an ANN. Like the synapses in a human brain, each link has the ability to send a signal to neighboring neurons. An artificial neuron can signal neurons that are connected to it after processing signals that are sent to it. The output of each neuron is calculated by some non-linear function of the sum of its inputs, and the "signal" at a connection is a real number. Edges refer to the connections. The weight of neurons often changes as learning progresses. In this study it gives an accuracy of 89%.

5.3. Stochastic Gradient Descent (SGD)

It is a simple yet highly effective method for fitting linear classifiers and repressors under convex loss functions, such as (linear) Support Vector Machines and Logistic Regression (Diab, 2019). SGD has been present in the machine learning field for a while, but in the context of large-scale learning, it has just recently attracted a lot of attention. Large-scale and sparse machine learning issues that arise frequently in text classification and natural language processing have been successfully tackled with SGD. The classifiers in this module are easily scalable to situations with more than 105 training examples and more than 105features because the data is sparse and for that we use it as our model for analyzing sentiments with an accuracy of 91% (Table 4).

Table 4. SGD Result

	Precision	Recall	F1-Score	Support
Negative	0.95	0.92	0.94	638
Neutral	0.86	0.94	0.90	639
Positive	0.94	0.87	0.90	643
Accuracy			0.91	1920

6. Discussion

Based on the results given by the models mentioned above we noticed that they achieved a high accuracy that pass 89% which is better than the accuracy of all most all the authors mentioned in the related work section despite the fact that they used different datasets and different approaches to perform the classification task.

7. Conclusion

This paper proposed a solution to analyze Arabic sentiments and classify them in three categories Positive, Negative and Neutral. This study was performed on a dataset consisting of 4800 tuples where we apply a simple treatment to filter, clean and balance it. The classification task was performed using three approaches namely Stochastic Gradient Descent (SGD), Naive Bayes and Neural networks where we witness a high accuracy that passed 90% for the first two. We look forward to define a method that can turn emojis to words to take them into consideration in the classification process.

References

Ali, M. M., 2021. Arabic sentiment analysis about online learning to mitigate covid-19. Journal of Intelligent Systems, 30(1), 524-540.

Diab, S., 2019. Optimizing stochastic gradient descent in text classification based on fine-tuning hyper-parameters approach. A case study on automatic classification of global terrorist attacks. arXiv preprint arXiv:1902.06542.

El-Masri, M., Altrabsheh, N., Mansour, H., & Ramsay, A. (2017). A web-based tool for Arabic sentiment analysis. Procedia Computer Science, 117, 38-45.

Folorunsho, D., Oct 13, 2020. Arabic sentiment analysis, in “Towards Data Science”. https://towardsdatascience.com/arabic-sentiment-analysis-5e21b77fb5ea.

Hadj ameur M.S., Aliane H., 2021. AraCOVID19-MFH: Arabic COVID-19 multi-label fake news & hate speech detection dataset. Procedia Computer Science 189, 232-241.

Loukas, S., Oct 12, 2020. Text classification using Naive Bayes: Theory a working example, in “Towards Data Science”. https://towardsdatascience.com/text-classification-using-naive-bayes-theory-a-working-example-2ef4b7eb7d5a.