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

A logistic regression algorithm for Arabic hate speech detection

Abdelmounim Sellidj

USTHB, BP 32 Bab-Ezzouar, 16111 - ALGER, ALGERIA
LINS-USTHB Laboratory of Instrumentation

Abstract

Arabic language is one of the most popular languages and it is widely used in social media networks. During the pandemics, the spread of fake news, rumors, hate speech and spams increased dramatically which makes the detection of the misinformation sources very important and very helpful to control the situation. A lot of Arabic natural language processing (ANLP) works are proposed in the literature to solve such problems, in this paper we propose a time efficient and high precision and accuracy algorithm for Arabic Hate speech detection.
A classical Machine Learning (ML) logistic regression algorithm is used in this ANLP work to detect hate speech, the data of this work are collected from Twitter social media during the COVID-19 pandemic, we use 80% of the data to train our algorithm and 20% of data to test it. The proposed algorithm has high accuracy and precision in the tested comments (a precision of 88.77% an accuracy of 98.48%). This work shows that, the classical ML algorithms have good performances in such problems.

Keywords: Machine learning; Natural Language Processing (NLP); Arabic NLP (ANLP); logistic regression, hate speech detection;

1. Introduction

The Arabic language is one of the most spoken languages in the world, a lot of Arabs used social media during the COVID-19 pandemic to express their feelings or share news, these were ideal circumstances to spread rumours, fake news, depressing phrases, and hate speech comments in social media. This misinformation affected the decision of people and made them resist and even oppose the recommendation of the concerned authorities (doctors, security agents …etc.), they refused to take treatment, vaccines, and

medicaments for COVID-19, and worse than that, they even refused to listen to the protection measurements Hadj Ameur et Aliane, 2021a, Khan et al., 2021. If we take into account all the problems that result from this misinformation and hoaxes then we must detect and stop the spread of misinformation.

A lot of works in the literature propose methods to solve this problem, they use several natural language processing (NLP) techniques using Machine Learning (ML) or Deep Learning (DL) algorithms to detect or classify sentiments, hate speech, rumors, hoaxes and fake news in the Arabic language Hadj Ameur et Aliane, 2021a, Khan et al., 2021, Amoudi et al., 2022, Jararweh et al., 2019 Guelil et al., 2021, Pirelli et Zerghili, 2017, Hadj Ameur et Aliane, 2021b.

Arabic could be written in three main varieties in social media, classic Arabic (CA), Modern Standard Arabic (MSA), or Arabic Dialect (AD), MSA and CA are written only in the Arabic language alphabet, but AD is written mostly in Latin script (Arabize). The proposed program can detect hate speech written only in the Arabic alphabet

In this work a machine learning program using natural language processing (NLP) techniques is presented which has the capability of detecting hate speech in the Arabic language on the social network Twitter. We used a time optimized Logistic Regression machine learning algorithm that was implemented in Octave Software. We note that, this program is not using any machine learning library or any pre-trained model. Also, in this work only the Arabic alphabet are token in consideration, all other languages characters and symbols are filtered, this can be a failure in our proposed approach if a non-Arabic hate speech is included in the comment, this model can't detect it.

2.Related works

Arabic language is one of the most complex languages in the world, due to its complexity in terms of meaning and structure, as an example for the complexity of its meaning the word "بلغ" has different meanings it could be "reach, arrived or tell", we can understand the specified meaning only if the word is diacritized or from the context, furthermore the structure can be more complex, one word which can consisted of the stem (principle) word with a suffix or prefix and also it can contain a gender or number indices (one person, two persons or plural) Kanan et al., 2019, example:

Table 1: Example of an Arabic word

Arabic term: لتنبؤونهم English meaning: To inform them				
Antefix	Prefix	Root	Suffix	Postfix
ل	ت	نبأ	ون	هم

AD (Arabic dialects) are complex because they don't have fixed rules. Because of these issues and other problems, ANLP needs a lot of efforts and research work to get a better treatment of the Arabic texts in social media, some previous research works are mentioned in Alzand et al.,2015. In recent years, a lot of works in this topic are presented in the literature, as mentioned in Guellil et al.,2021, some of the previous works in ANLP deals with Classical Arabic (CA), Modern Standard Arabic (MSA), Arabic Dialect (AD) or Arabizi. Also, we remark that some works use predefine libraries to get better performances and high accuracy like Hadj Ameur et Aliane.,2021a. recent review work Guellil et al., 2021, Larabi et al., 2018, Alruily, 2021, Kannan et al., 2019 and Zahidi et al., 2019 summaries the recent works, detail the different techniques used in ANLP, present the challenges of this domain and propose some future research axes.

Some works dealt with the Arabic language vocabulary and grammar such as Larabi et al., 2018, there objective was to simplify the treatment of Arabic texts which has different rules compared to English language (no capital letter for non, the pronouns could be either separate or attached to the principal word, one

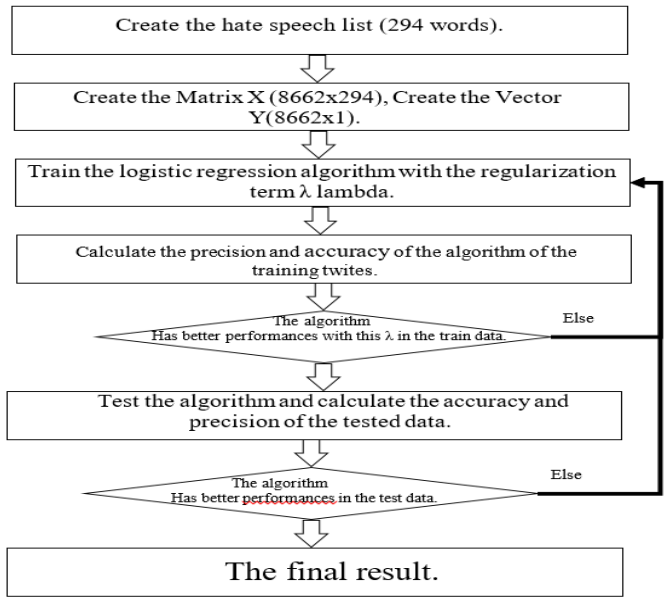

Figure1. proposed method

word could have a lot of synonyms, the letters have different forms according to its position in the word and other vocabulary or grammar differences). Jararweh et al., 2019, Guelil et al., 2021, Pirelli et Zerghili, 2017, Murzsi et al., 2022, Abu shquir, 2019, Larabi et al., 2018, Alruily, 2021, Kannan et al., 2019 and Zahidi et al., 2019, Bahurmuz et al., 2022, all treat the ANLP in difference manners, they cover different aspects of the domain (reviewing, data processing using pre-trained libraries, data processing according to the nature of Arabic language …etc.);

3. Proposed Algorithm and data processing

In our work we use the dataset from Hadj Ameur et Aliane, 2021a to detect hate speech from 10828 tweet on Twitter regarding the COVID-19 pandemic. We use 8662 comments to train the LR algorithm and 2166 comments to test our trained LR algorithm. For the Software and Hardware environments we use GNU Octave as a programming language to execute our program in an HP laptop with an I3 two cores CPU @2.4GHz and a DDR3 RAM of 16GO in windows 10.We followed the algorithm steps as mentioned in the algorithm diagram (Figure 1). For ,data preprocessing, we filtered all non-Arabic characters, numbers, URLs and emojis which allows us to treat only the Arabic tweet texts because we are interested only in Arabic words, this has several advantages: an important reduction in time execution for both training and testing of the model, much less complex model and the model is trained well for Arabic texts. However, this can also be a source of problems: we can mention as an example the model can't detect tweets with hate speech in another language. Note: Our pre-processing doesn't differentiate between CA and MSA neither between the same words with a suffix or a prefix as example "كورونا" and "الكورونا" in (Table 2) is the same word ("Corona" and "the Corona") but this program considers it as different words, see the algorithm diagram, Figure 1.

3.1 Selecting the algorithm

After preprocessing, we select the suitable ML algorithm to detect if a tweet contains hate speech or not. In reality, all the classic ML algorithm could be used to detect hate speech and they have good accuracies according to Mercan et al.,2021, which presents a comparative study for hate speech and offensive language detection from social media using classical ML algorithms such as "logistic regression LR, random forest RF, naive Bayes NB, and support vector machine SVM" and advanced deep learning-based models "recurrent neural networks RNN and bidirectional encoder representations BERT" and a comparison between the efficiency of them, this study concludes that deep learning models show a slight high accuracy than the classical ML models about 3% between SVM 84.66% and BERT 87.78% (The best of each approach). In our case we choose the LR algorithm with regularization term and a gradient descend optimization algorithm because of it is efficiency for problems which have only two probabilities Kleinbaum et al., 2002, as in our case we detect if this is a hate speech comment or not.

3.2 List of hate speech words

After choosing the appropriate algorithm which fit with our application, we select a list of hate speech words from the training comments, for that we select the words which appear 10 times or more in all the training hate speech comments and we get a list of 294 word, to avoid the overfitting problem and to reduce the time of execution, since there are some words that are repeated many times in the training hate speech comments and they have an important role to detect hate speech comments so they have high weight so the words appears less than 10 times will have less effect in the algorithm decision (Table 2).

Table 2 The 10 highest detected words in the hate speech comments

N°	Appearance	Words	N°	Appearance	Words
1	844	الله	6	192	في
2	800	كورونا	7	191	الكورونا
3	733	يلعن	8	170	الصين
4	321	من	9	163	بيقهيت
5	310	و	10	140	اللي

3.3 Time optimization

In order to execute our program rapidly and effectively, first of all we filter all the training comments from non-Arabic characters, next step is to compare every training comment with our predefined hate speech library and we get a features matrix (X) of 8662 x 294 dimensions composed only of ones and zeros signifying what words in the hate speech library appears in the comment and this is valid for the comment be either a hate speech tweet or not, in the same time we create a vector (Y) of ones and zeros signify either this is a hate speech tweet or not, so we deal with binary numbers instead of strings, finally, we save the matrix X and the vector Y in a binary data format to use them later, all this steps help us to reduce and optimize the execution time and also it makes the saving and reading processes very rapid (ex: we reduced the file opening time by 3 minutes compared to other formats).

The, we train our model and we get the corresponding parameters (Theta) (this data could not be saved in a binary format). The next step is to test our model, for that we take all the test comments and filter them as we do for hate speech list, to create a new file with only Arabic words to reduce and optimize the execution time. We also execute each part separately to avoid the repeated work.

All parts of the program are executed in about 1h 16min 32.55s and we reduce the overall execution time by about 7min.

3.4 Algorithm performance and limitations

The algorithm has a precision of 88.77% an accuracy of 98.48% a recall of 95% and F1 score of 91.82% the system works very well for the proposed test comments, but for comments use English words or Arabize hate speech this program can't detect it, also if the tweet contains hate speech words or characters who doesn't belong to the list of hate speech word than the program fails.

We note that the number of the training comments is small (8661) specific to hate speech comments consists only of 985 with some repetitive comments or with the same format. This is also true for the testing comments, we have only 196 hateful comments and 1919 not hateful comments a total of 2116 test comments which makes the previous performances of this algorithm not very precise.

3.5 Some examples of the final results

Table 3. Examples of all possible cases of the classification program (1 and 3 right classification, 2 and 4 wrong classification)

Comments	Hatful	Detected as Hateful speech
1- @SaudiNews50 يارب يروح الفيروس يارب ☹️💔☹️☹️	No	No
2-كورونا له علاج أو لقاح والحوثي لا علاج له	Yes	No
3- ☹️ الله يلعن الكورونا و يلعن الصين	Yes	Yes
4- الله يلعن الكورونا حرمتنا من ممارسة كرة القدم ☹️ - واحد رجليه الثنتين يسار	No	Yes

4. Conclusion and future work

In this work, we treated 10828 comments in Arabic language in which 8661 comments (80%) of them used as training examples to train our logistic regression algorithm and 2116 comments (20%) of them used for test the performances of our proposed program, we used the classical Logistic Regression algorithm without any other libraries or pretrained models. Our algorithm fit well to this particular solution and it has an impressive performance a precision of 88.77% an accuracy of 98.48% a recall of 95% and F1 score of 91.82%.

Nevertheless, it has its own limitation:

- It detects hate speech only in MSA which use only Arabic alphabet in the written words.
- A small list of hate speech words compared to the pretrained libraries which can lead the algorithm to fail if a word how doesn't exist in the hate speech list is used in a new comment.

In this work we show that even the classical Algorithms of ML can solve NLP problems with high accuracy and precision performances. As future work, we plan to:

- Use more data to test the performances of this program.
- Compare the logistic regression ML algorithm with other classical ML algorithms such as Neural Network, Support Vector Machine.
- Compare this work with works deep learning and pretrained libraries.

References

Alansary, S., Nagi, M., Adly, N., 2013. A suite of tools for Arabic natural language processing: A UNL approach. In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)* (pp. 1-6). IEEE.

Alruily, M., 2021. Classification of arabic tweets: A review. *Electronics*, *10*(10), 1143.

Alzand, A. A., Ibrahim, R., 2015. Diacritics of Arabic Natural Language Processing (ANLP) and its quality assessment. In *2015 International Conference on Industrial Engineering and Operations Management (IEOM)* (pp. 1-5). IEEE.

Amoudi, G., Albalawi, R., Baothman, F., Jamal, A., Alghamdi, H., Alhothali, A., 2022. Arabic rumor detection: A comparative study. *Alexandria Engineering Journal*, *61*(12), 12511-12523.

Bahurmuz, N. O., Amoudi, G. A., Baothman, F. A., Jamal, A. T., Alghamdi, H. S., Alhothali, A. M., 2022. Arabic Rumor Detection Using Contextual Deep Bidirectional Language Modeling. *IEEE Access*, *10*, 114907-114918.

Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., & Nouvel, D., 2021. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, *33*(5), 497-507.

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

Hadj Ameur, M. S., & Aliane, H., 2021b. Aracovid19-ssd: Arabic covid-19 sentiment and sarcasm detection dataset. *arXiv preprint arXiv:2110.01948*.

Jararweh, Y., Al-Ayyoub, M., & Benkhelifa, E., 2019. Advanced Arabic natural language processing (ANLP) and its applications: introduction to the special issue. *Information Processing & Management*, *56*(2), 259-261.

Kanan, T., Sadaqa, O., Aldajeh, A., Alshwabka, H., AlZu'bi, S., Elbes, M., ... Alia, M. A., 2019. A review of natural language processing and machine learning tools used to analyze arabic social media. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)* (pp. 622-628). IEEE.

Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., Iqbal, A., 2021. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, *4*, 100032.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., Klein, M., 2008. Logistic regression. *A Self-Learning Tekst*.

Marie-Sainte, S. L., Alalyani, N., Alotaibi, S., Ghouzali, S., Abunadi, I., 2018. Arabic natural language processing and machine learning-based systems. *IEEE Access*, *7*, 7011-7020.

Mercan, V., Jamil, A., Hameed, A. A., Magsi, I. A., Bazai, S., Shah, S. A., 2021. Hate Speech and Offensive Language Detection from Social Media. In *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)* (pp. 1-5). IEEE.

Mursi, K. T., Alahmadi, M. D., Alsubaei, F. S., Alghamdi, A. S., 2022. Detecting islamic radicalism arabic tweets using natural language processing. *IEEE Access*, *10*, 72526-72534.

Pirrelli, V., & Zarghili, A., 2017. Arabic Natural Language Processing. *Journal of King Saud University-Computer and Information Sciences*, *29*(2), A1-A3.

Shquier, M. M. A., 2019. Novel Prototype for Handling Arabic Natural Language Processing: Smart Morphological Analyser. In *2019 Second International Conference on Artificial Intelligence for Industries (AI4I)* (pp. 1-8). IEEE.

Zahidi, Y., El Younoussi, Y., & Azroumahli, C., 2019. Comparative study of the most useful Arabic-supporting natural language processing and deep learning libraries. In *2019 5th International Conference on Optimization and Applications (ICOA)* (pp. 1-10). IEEE.