Report 1

May 8, 2023

1 Motivation

Social media platforms have become an integral part of our daily lives, where users express their opinions and share their experiences on a wide range of topics. With the massive volume of data generated on these platforms, it's becoming increasingly challenging to manually classify and analyze tweets based on their topics. This is particularly true for Arabic tweets, which present unique linguistic challenges and require specialized tools and techniques.

That's where our NLP project comes in! By leveraging the latest advancements in NLP and machine learning, we aim to develop an accurate and efficient model that can classify Arabic tweets into different categories based on their topics. This will enable individuals and organizations to gain valuable insights into what people are talking about and make informed decisions based on real-time data.

Our project has the potential to make a significant impact on various fields, such as marketing, politics, and social sciences. With the increasing importance of Arabic as a global language, our project also has the potential to contribute to the development of Arabic language processing tools and resources.

We will fine-tune a pretrained model MARBert model. The network architecture is the same as BERTBase (12 layers, 768 hidden units, 12 heads).

2 Related Work

There have been several related works on Arabic tweet classification that may be relevant to this project. One of them is this paper which proposes a model for Arabic sentiment analysis using a Twitter dataset and deep learning models with Arabic word embedding. It uses the supervised deep learning algorithms on the proposed dataset. The dataset contains 51,000 tweets.The experiment has been carried out by applying the deep learning models, Convolutional Neural Network and Long Short-Term Memory while comparing the results of different machine learning techniques such as Naive Bayes and Support Vector Machine. The accuracy of the AraBERT model is 0.92

3 Dataset

In our project we used SANAD dataset. SANAD Dataset is a large collection of Arabic news articles. The articles were collected using Python scripts written specifically for three popular news websites: AlKhaleej, AlArabiya and Akhbarona. SANAD has seven categories, which are: Culture, Finance, Medical, Politics, Religion, Sports and Technology.For this project, we used a balanced subset of the dataset in which each category contains 6,500 articles, so we have in total 45,500 articles.

4 Challenges

The limited availability of Arabic NLP resources is a significant challenge in developing NLP models for Arabic text classification, including tweets. Unlike English, Arabic is a morphologically rich language with a complex grammar and unique linguistic features, such as the presence of diacritics and the right-to-left script.

Collecting and preprocessing a large and diverse Arabic text dataset can help address the limited availability of NLP resources. Preprocessing techniques such as stemming, normalization, and stop word removal can also help to simplify the text and reduce the complexity of the language.

5 Preprocessing Steps

Preprocessing the data involves the following steps

1. **Remove stop words** : For this task, we used the Arabic stop words corpus that is provided by NLTK. However, this list of Arabic stop words provided by NLTK was not comprehensive, so we added an extra stop words using an external file contains more than Arabic 2000 stop words.
2. **Remove English letters**.
3. **Remove numeric digits** : for this task, we take into consideration both of the Arabic and English digits to be removed.
4. **Remove Punctuations** : For this task, we used the constant punctuations that is provided by string module in python. This constant contains all ASCII punctuations characters like "!", "?", etc.. However, this set was not sufficient as the Arabic contains more different punctuations such as the Arabic punctuation "", so we add manually a customized set of extra punctuations to be more comprehensive.
5. **Remove extra spaces**.
6. **Remove Arabic diacritics** : also known as tashkeel such as Fatha, Kasra, Damma and so on.
7. **Normalize arabic letter** : this involves replacing different forms of the same letter with a standard form. This can improve the accuracy of text analysis and classification by reducing the number of unique characters that the algorithm has to consider.
8. **Normalize words** : This is done by applying stemming, and we used for this ISRIStemmer algorithm from NLTK.

6 Data Analysis

First, we checked if out data contains null values and we found it does not. However, we found that we have 15 duplicate values (rows) in the dataset, so we removed them. Before removing duplicates, the date was balanced, 6,500 articles for each category. But, because that the number of removed rows is very small, the data is still considered as balanced as shown in Figure 1.

Also we used word cloud to figure out what is the common words in each category, There aresome examples in the end of the report.

7 References

1. Sanad dataset : https://data.mendeley.com/datasets/57zpx667y9

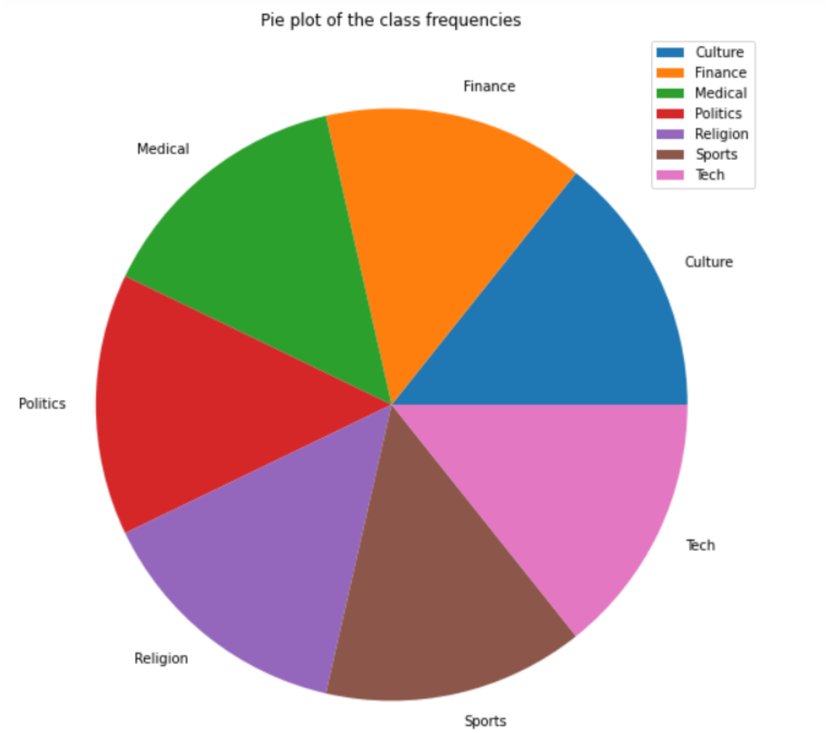

Figure 1: This frog was uploaded via the file-tree menu.

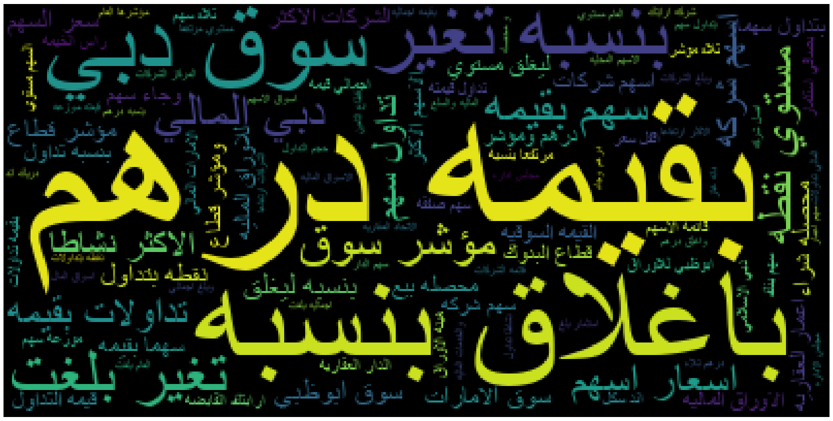

Figure 2: Finance word cloud.

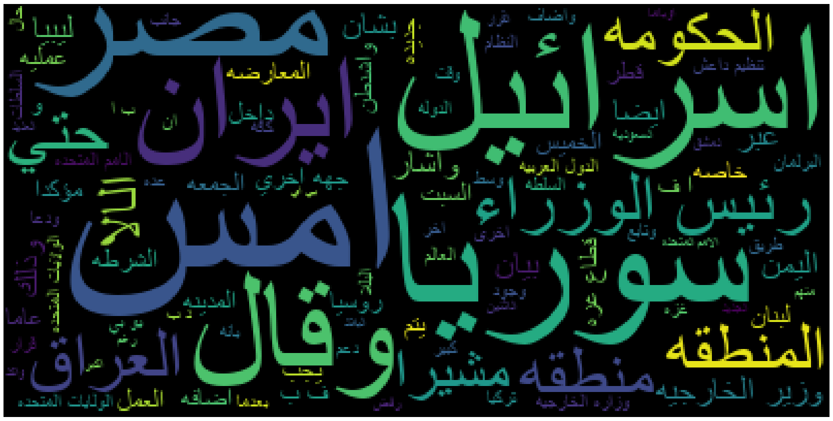

Figure 3: Politics word cloud.