Project Final Report Outline

PROJ 201 Project Final Report

Project Title: Searching most accurate and existing methods to detect the language of texts

Name, Surname & ID of group members: Osman Enes Erdoğdu & 29344, Sinem Gül Kaya & 30298

Supervised by: Yücel Saygın

Month, day, year: 01.06.2023

Abstract

The main goal of this project is to find which model is the best for detecting English text by using NLP tools. The best method can predict better if a text is English or non-English. To be able to detect which model is better for detecting English texts first we will search for some common existing tools to predict language. Then we need to find the dataset to test the tools and evaluate and compare the tools. After that, we will test the best model. Finally, from our results, which are in the appendix part, we must answer whatever our main goal is. For us, that main goal or question is whether the text we are working on is English or non-English. In the remainder of our final report, we will explain the steps that we worked on and clarify the results that we found while testing the data. At the end of our report, based on the results that we reached, we will examine whether it is possible to detect the language of YouTube Video by title or description.

Introduction

NLP means natural language processing. In this Project, our aim is to test some NLP tools and analyze them so we can find which model is the best for detecting English texts. In this project, we tested a Python-based solution for detecting English texts. To be able to detect the language of the texts we must remove some problematic characters such as emojis etc. Which we will explain in the methods and materials part. We need to do that because punctuation mistakes, emojis different characters, etc. obstruct the code to detect the language of the text. There are many videos on YouTube and some of them are English, some of them are non-English, and some of them are mixed. Sometimes it can be difficult to predict which language is used in the content of the video because different languages can be used in the description or title. So, the main question we want to consider is whether we can predict the language of the YouTube video (English or non-English) based on the title or description. And which one has the higher possibility to be accurate? We followed some steps and reached a conclusion that we will detailly explain in the other parts.

Methods & Materials

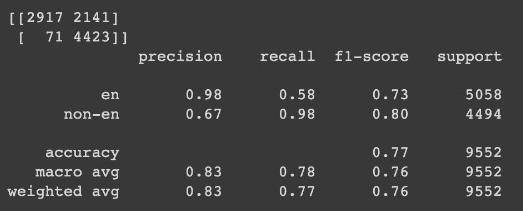
We followed specific steps in this project where we investigated whether we could predict the language of the YouTube video (English or non-English) based on the description or the title. Firstly, we searched the existing methods for detecting English and non-English texts. Then to decide the most accurate one, by using an algorithm we collected the data from Twitter and created our dataset. Our data contains 10502 tweets. But we had to delete problematic characters from the database to clean up the tweets in our preprocessing step. First, we removed emojis (that contain emoticons, symbols & pictographs, transport & map symbols, flags, Chinese characters, and dingbats…), tags, hashtags, links, new lines, e-mail addresses, punctuation marks, and lastly extra spaces. After removing the problematic characters, 9552 tweets remain. 5058 of them are in English and 4494 of them are non-English. Then, we installed 3 methods we found to detect English texts that are Langid, Langdetect, Pylcd2. We used a confusion matrix to interpret the results of our model and to examine the differences in the relationship between actual and predicted values. We decided to the best model while comparing the accuracy rate which can be seen in the results part (A., B., and C.). After deciding on our best model, we tested the YouTube videos we collected on this program. Our data includes many videos in branches of Sociology, Public Relations, Psychology, Politics, English Language and Literature. Our program first predicted which language the video was in based on the title of the YouTube video and then the description of the video. Based on these results, we created a confusion matrix and argued that based on accuracy rates, whether the language of the video can be predicted by looking at its title or description.

Results

In this part of our project, we will explain the results we obtained in the steps we mentioned in the methods & materials section. First of all, we compared the data in the confusion matrix and checked the accuracy rates. While looking at these rates, we decided that Langid was the best among the 3 models with an accuracy rate of 84% (can be seen in part B.). After that, we used this model on our code to predict if we can decide on the language of the video based on looking at the title or description. We ran our code, and we reached some conclusions. We created two confusion matrixes (that are in part D. and E.). Our model predicts the language of the video by looking at the title with a 53% accuracy rate, and by looking at the description with a 57% accuracy rate. Thus, by looking at these results, we cannot conclude in which language a video is based on the title and description. However, when we predict the language of a video by looking at the description of a video, considering the accuracy rates, the probability of our prediction being correct is much possible than the probability of being correct when we predict by looking at the title of the video.

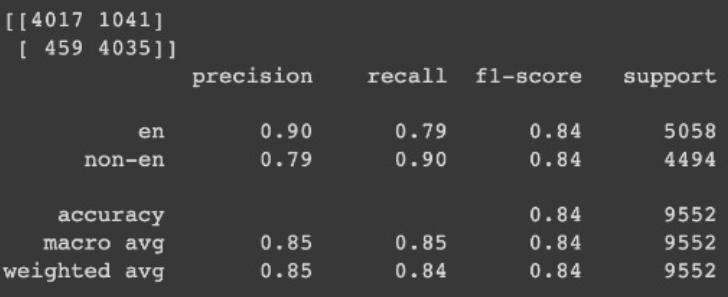
A.

Pycld2



B.

Langid

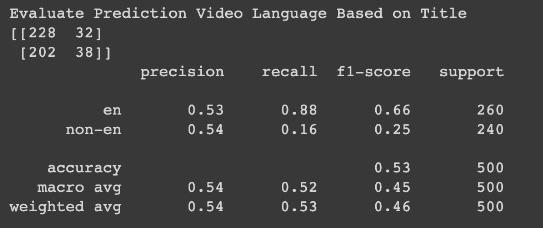


C.

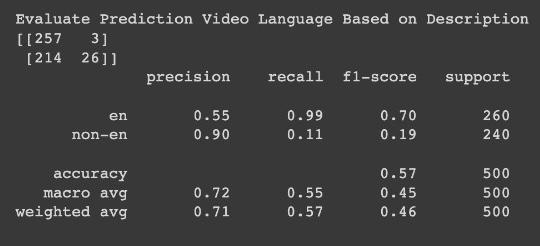
Langdetect



D.



E.



Discussion and Conclusion

So, in a conclusion, we followed the five steps to achieve our goal which was detecting the language of the text. First, we searched for common tools. Then we find a dataset (which was descriptions of YouTube videos in our case). We evaluate and compare the best tool and tested them on our data. And finally, achieve our goal and detect the language with the most accurate tool. This project is really important and has great potential for the future. Machine learning is an important topic in science. And through the project, we learned so much. In the beginning, we learned what NLP and machine learning are then we started to examine and learn about the code that helps our goal. This project is important because as technology gets into our lives more and more machine learning becomes more important. For example, in our project, in the future, most of the companies who want to collect data from the internet can use this type of code that we worked on in this project. Also, Artificial Intelligence getting better and better so our project can be used in a lot of areas in the future which is why it’s very important.