ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
ΠΜΣ "Πληροφοριακά Συστήματα & Υπηρεσίες"

SPECIALIZATION: Big Data & Analytics

COURSE: Data Mining and Preparation

TEACHER: Maria Chalkidis

POSTGRADUATE INTERNSHIP FEB. 2022-23

Academic year: 2022-23

Students:

Magirias Georgios, A.M.: me2220

09/02/2023

Introduction

It is a fact that nowadays, the phenomenon of fake news is an issue that concerns our society to a very large extent. Although it is not something we come across first time in this day and age, but we can trace it back over the years, it is commonly accepted that in recent years, due to the development of technology, it has take on even greater dimensions. So much information that it receives person on a daily basis from a variety of sources, as well as his weakness many times to process them, due to their large volume, has led to difficulty their recognition. For this reason, an effort is made towards finding ways their recognition, so that this phenomenon is eliminated as best as possible degree.

But what do we mean by "fake news" or better by the more widely known term "fake news"? Fake news is fabricated news, which is either not do not correspond to reality at all or are based on real events, but including false information in order to mislead the receiver (Schlesinger, 2017).

According to the categorization made by Wardle & Derakhshan (2017), there are three types of false information, those that are false and created to cause harm to an individual or group of individuals, those which are false, but they do not intend to harm anyone and finally those based on reality and have the intention of using them to cause harm to someone individual, social group or organization.

The motivations for spreading fake news can be traced to two main ones sectors, in the economic and the ideological. Regarding the financial, the goal is through of disseminating news that will be attractive to the public to arise financially income, while in the ideological field, in this the goal is promotion specific ideas (Tandoc, Lim & Ling, 2018).

Whatever the motivation behind the spread of fake news, this phenomenon must to be eliminated immediately, so as to provide the receivers with timely information, h

which will not include false news that only aims to mislead public for various reasons.

Literature review

Researchers in recent years have made a continuous effort to construct algorithms which with the necessary training will be able to identify with success fake news. Great importance for its effectiveness algorithm has how up-to-date the training data will be. Ajit Patil, Rajeshri Kalwale, Harshita Kaushik and Pranita Thorawase (2022) use the technique web scraping to retrieve the training data targeting with this way to timeliness and reliability. Then through PCA they managed to reduce the dimensions and finally used categorization methods such as Logistic Regression, Decision Trees and K Nearest Neighbors (K-Nearest Neighbor) where they managed to achieve high accuracy rates.

P. Nair and I. Kashyap (2019), concluded that by applying techniques resampling and intervening on outliers during its stage preprocessing, they manage and normalize their data and succeed greater accuracy in classification.

MF Mridha, AJ Keya, MA Hamid, MM Monowar and MS Rahman (2021) they conclude that Deep Learning techniques achieve greater accuracy in prediction compared to traditional Machine Learning methods (Neural networks) and despite the fact that they occupy a larger percentage of memory, however they excel in execution times.

Rusli A., JC Young and NMS Iswari (2020) categorizing with and without "stemming" and removing "stop-words" concluded that in the second case they achieved a better F-score by 0.02, but this benefit was not enough to to cover the cost in time required to carry out the above methods.

The objective of this paper was to create and evaluate two machine learning models, a Logistic Regression model and a Support model

Vector Classifier (SVC), to classify news as reliable or unreliable. At
then used a dataset from Kaggle, "Fake News", to
models are trained and tested. After evaluating the performance of the models,
created an application that allows users to enter news articles
and receive predictions from the trained models.

Methodology

Data Description and Preprocessing

The dataset used in the work was Fake News from the Kaggle website. Each line of the data corresponds to an article, for which its author and its title are known. Also in the data set there is an additional feature that categorizes each news item as reliable or unreliable. The data contains 20800 observations. Below are the features that comprise the data, followed by the first five observations of the set.

Fake News dataset	
id	Unique id for a new article
title	The title of a news article
author	Author of the news article
text	The text of the article
label	1=Unreliable, 0= Reliable

	id	title	author	text	label
0	0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

Initially the data was loaded into python using the pandas library and checked for null values. The null values were replaced with an empty character and then a new "content" attribute was created, which contained the information of all the original attributes except "id" and "label".

The data contained in the new "content" attribute undergoes the following preprocessing:

1. Text content is cleaned by removing any non-letter characters (az or AZ) and replacing them with a space character.

2. The cleaned text is then converted to lowercase.
3. The cleaned text is split into individual words.
4. The words are then modified using the PorterStemmer object, which is a tool for reducing words to their base or root.

5. Stopwords are removed from the word list.
6. Finally, the processed words are joined together again into a single string.

After the above pre-processing was done, then using the TfidfVectorizer of the scikit-learn library, the text was transformed into numerical data, representing the importance of the words of each document, taking into account the frequency with which they appear in the set of documents. The data was then split into training and control data, with an 80/20 ratio.

Training Classifiers

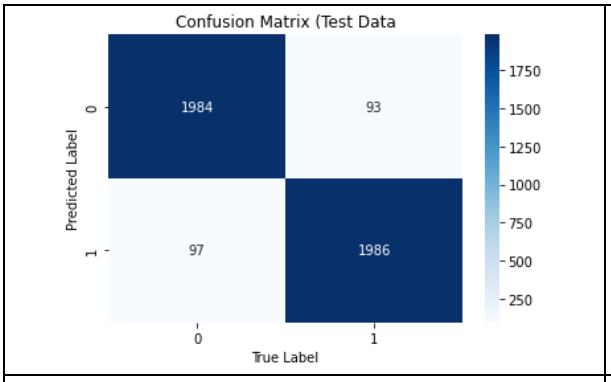
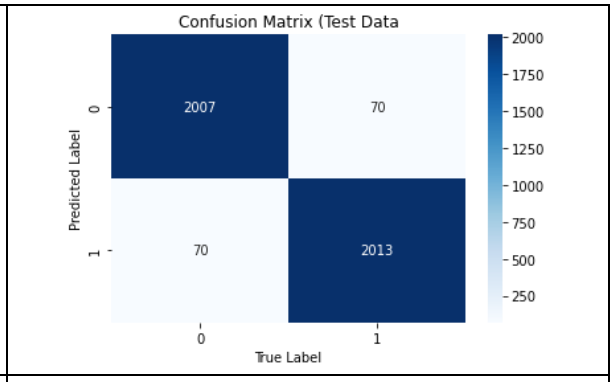
With the data obtained from the preprocessing, the training of two different classifiers, Logistic Regression and Support Vector Machines (SVC), of the scikit-learn library was performed. The training times of the models were also measured. Finally, it should be mentioned that the resulting trained models were saved using the joblib library, so that they could be used in the context of creating the application.

Results

The accuracy score was used as a measure to evaluate the performance of the classifiers, which was calculated for both the training data and the control data. The following table shows the training times of the models, as well as the accuracy they achieve, while the time needed for each model to predict the control data is also calculated.

	Training time (seconds)	Train Accuracy (%)	Prediction time on test set (sec)	Test Accuracy (%)
Logistic Regression	1.44	98	0.007	95
SVC	689.91	99.8	58.74	96.6

Then the confusion matrices are presented, resulting from the predictions of the two models on the control data.

	
Logistic Regression	Support Vector Classifier

Conclusions

Observing the results, the following conclusions are drawn:

- The training time of the modelSVC is about 500 times longer than the time taken by the Logistic Regression model.
- Both models achieve high accuracy on the training data and control, with the SVC model achieving slightly better accuracy.
- The modelSVC was much more time-consuming than the Logistic Regression model in the process of predicting control data

Bibliographical references

Mridha, MF, Keya, AJ, Hamid, MA, Monowar, MM & Rahman, MS (2021). A Comprehensive Review on Fake News Detection With Deep Learning, IEEE Access, vol. 9, pp. 156151-156170, 2021, doi: 10.1109/ACCESS.2021.3129329.

Nair, P. & Kashyap, I. (2019, February). Hybrid Pre-processing Technique for Handling Imbalanced Data and Detecting Outliers for KNN Classifier, International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 460-464.

Patil, A., Kalwale, R., Kaushik, H., & Thorawase, P. (2022). Fake News Detection Using Machines Learning Algorithms. Gis Science Journal

Rusli A., JC Young and NMS Iswari, "Identifying fake news in Indonesian via supervised binary text classification", Proc. IEEE Int. Conf. Ind. 4.0 Artif. Intell. Commun. Technol. (IAICT), pp. 86-90, Jul. 2020.

Schlesinger, R. (2017). Fake News in Reality. Retrieved from

https://www.usnews.com/opinion/thomas-jefferson-street/articles/2017-04-14/what-is-fake-news-maybe-not-what-you-think

Tandoc Jr, EC, Lim, ZW, & Ling, R. (2018). Defining “fake news” A typology of scholarly definitions. Digital journalism, 6(2), 137-153

Wardle, C., & Derakhshan, H. (2017). Information Disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe Report, 27