Machine Learning to Identify Fake News for COVID-19

Marianna ISAAKIDOU[a,1], Emmanouil ZOULIAS[b] and Marianna DIOMIDOUS[a]

[a] *Faculty of Nursing, National and Kapodistrian University of Athens, Athens, Greece*
[b] *Health Informatics Laboratory, Faculty of Nursing, National and Kapodistrian University of Athens, Athens Greece*

Abstract. International Organizations are seriously concerned about the fake news phenomenon. UNESCO has defined the term of misinformation/disinformation, which are the two faces of fake news. European Commission has conducted a survey about "Fake News" through EU citizens to estimate the awareness and people behaviour related to the appearance of fake news and disinformation on electronic. The findings are quite worrying, since about 40% come across fake news daily and 85% evaluate fake news as a problem. The aim of this work is to introduce an Artificial Intelligence approach, the Decision Trees algorithm to identify fake news on the COVID-19.

Keywords. Fake News, COVID-19, Machine Learning, Artificial Intelligence, Natural Language Process.

1. Introduction

The rise of fake news highlights the erosion of long-standing institutional bulwarks against misinformation in the internet era and the concern over the problem is global. However, much remains unknown regarding the vulnerabilities of individuals, institutions, and society to manipulations by malicious persons. A new system for defence is needed [1]. According to UNESCO's definition, misinformation/disinformation is the false information that is spread under the guise of news, which in fact is very often impressively written, to arouse people's interest [2]. A survey about "Fake news" was conducted online by the European Commission in February 2018, the aim was to investigate EU citizens' awareness and attitudes towards the phenomenon of fake news and disinformation existence. Only a few respondents have the position that they 'totally trust' all available sources. A rather high percentage of 37% respondents believe that they come across fake news daily or every other day, and additionally 31% of the responders' state that come across fake news at least once a week. A very high percentage of 85% respondents think that fake news is a serious and active problem in their country [3].

Healthcare specialists encounter, apart from the urgent need for rapid and adaptable knowledge to handle COVID-19 disease, also encounter the enormous bulk of new information spreading as fake news about health issues. This phenomenon

---

[1] Corresponding Author, Marianna Isaakidou, PhD cand. Faculty of Nursing, School of Health Sciences, National and Kapodistrian University of Athens, Athens, Greece; E-mail: marianis@nurs.uoa.gr.

particularly affects the public health policy effectiveness, especially during pandemics like COVID-19 [4]. Artificial Intelligence can support health by analysing massive amount of health data to reveal hidden information [5]. Against COVID-19 pandemic seven major types of significant applications of AI against COVID-19 pandemic focused on health monitoring, treatment, contact tracing, and drugs-vaccines [6].

Therefore, a decision support system (DSS) that discriminates fake and real news provide decision support capabilities to experts for discriminating any new information, as well for the public. A DSS could be a modern tool to confront the Infodemic "virus" during any pandemic. Pattern recognition techniques have been widely used towards the development and application of decision support systems to aid diagnosis [7]. These techniques include Decision Trees [8], Support Vector Machines [9], Artificial Neural Networks (ANNs) [10], k-NN [11]. There is a serious gap regarding solutions for reducing fake news on public health issues and particularly on the COVID-19 pandemic. The scope of this work is to introduce a decision support system (DSS) for the classification of fake and real news.

2. Methods

2.1. Data Description and Preparation

The data used on this work are by the multimodal repository of ReCOVery that facilitates reliability assessment of news on the COVID-19. Xinyi Zhou et al search and investigate 2304 new articles on coronavirus made by publishers of various political polarizations from various countries worldwide and retrieved by22 reliable and 38 unreliable websites. The data include various information like url, author, title, image, body_text, country, and other. In our work we explore title and body text [12].

The pre-processing procedure from the standard text analysis of the Natural Language Processing (NLP) using Python code based on "scikit-learn" Python library [13]. The text analysis we followed steps that are referred in the literature [14]. The first step is to shuffle the data to prevent bias. Then follows the step of lowercase all words, removal of punctuation. The next two steps filtering of "Stop Words" to exclude common words and tokenization with white space are from the Natural Language Toolkit (nltk) [15].

2.2. Classification Procedure

The classification procedure based on text mining analysis using Python code based mainly on «scikit-learn» Python library [13]. The modelling process will consist of vectorizing the "text_body" and "title" columns of the data, then applying Term Frequency-Inverse Document Frequence (TF-IDF) [16]. TF-IDF weight is a statistical measure used to evaluate how important a word is to a document in a collection or full text. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the text. The normalized Term Frequency (TF) computes the number of times a word appears in a document, divided by the total number of words in the same document. The Inverse Document Frequency (IDF), computed the logarithm of the number of the documents in the text divided by the number of documents where the specific term appears. Finally, a classification machine learning algorithm.

2.3. Decision Tree Classification

There are various top-down DTs inducers like CART [8]. The scikit-learn Python library uses an optimised version of the CART algorithm. The "DecisionTreeClassifier" in python is a classifier, that like other classifiers, takes as input of data parts, the training samples, and the testing samples. A randomised method used for splitting the data into train and test sets, and the proportion of the dataset included in the train split is 20% of total data. In addition to that the python algorithm parameter random_state has the value 42 as a seed to shuffle the data before applying the spliton. In this work we used various combination of parameter values specific for the Decision Tree. Particularly, used the Criterion, which is the function to measure the quality of a split. The proposed supported criteria are for the Gini impurity the "gini" value and for the information gain the "entropy" value. The second parameter is the Splitter, which is the strategy used to choose the split method take place at nodes. Available strategies are best split, using "best" value, and the best random split, using "random" value. Finally, the Max Depth of the Tree parameter, which is the number of that the Tree nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. In this work tested various combination (Table 1) of parameters values on the available data.

Table 1. Combinations of Decision Trees Parameters

Criterion	Splitter	Max Depth
Entropy (for the information gain)	Best	20
Gini (Gini impurity)	Best	20
Entropy (for the information gain)	Random	20
Gini (Gini impurity)	Random	20
Entropy (for the information gain)	Best	40
Gini (Gini impurity)	Best	40
Entropy (for the information gain)	Random	40
Gini (Gini impurity)	Random	40

3. Results

The algorithm applied separately on titles data as well as on main body text data. The results of the various tests appeared in Table 2, as overall accuracy (fraction of true negatives and true positives) for the various parameter values combinations.

Table 2. Percentage of Averaged Classification Accuracy - 10 Independent Algorithm Iterations

Parameters Combination	Overall Accuracy Body	Overall Accuracy Title
Entropy (for the information gain) - Best - 20	98,09%	72,09%
Gini (Gini impurity) - Best -20	96,05%	72,09%
Entropy (for the information gain) - Random - 20	95,53%	72,09%
Gini (Gini impurity) - Random - 20	94,01%	72,09%
Entropy (for the information gain) - Best - 40	100%	72,09%
Gini (Gini impurity) - Best - 40	99,13%	72,09%
Entropy (for the information gain) - Random - 40	99,87%	72,09%
Gini (Gini impurity) - Random - 40	96,44%	72,09%

The Confusion matrices for the best and worst cases are illustrated in Figure 1. It is worthwhile to mention that the Decision Tree algorithm has its worst results, regardless

the parameters used, when applied on the Title document. The title of every new is mainly the shortest part and the algorithm has no adequate words to be trained properly.
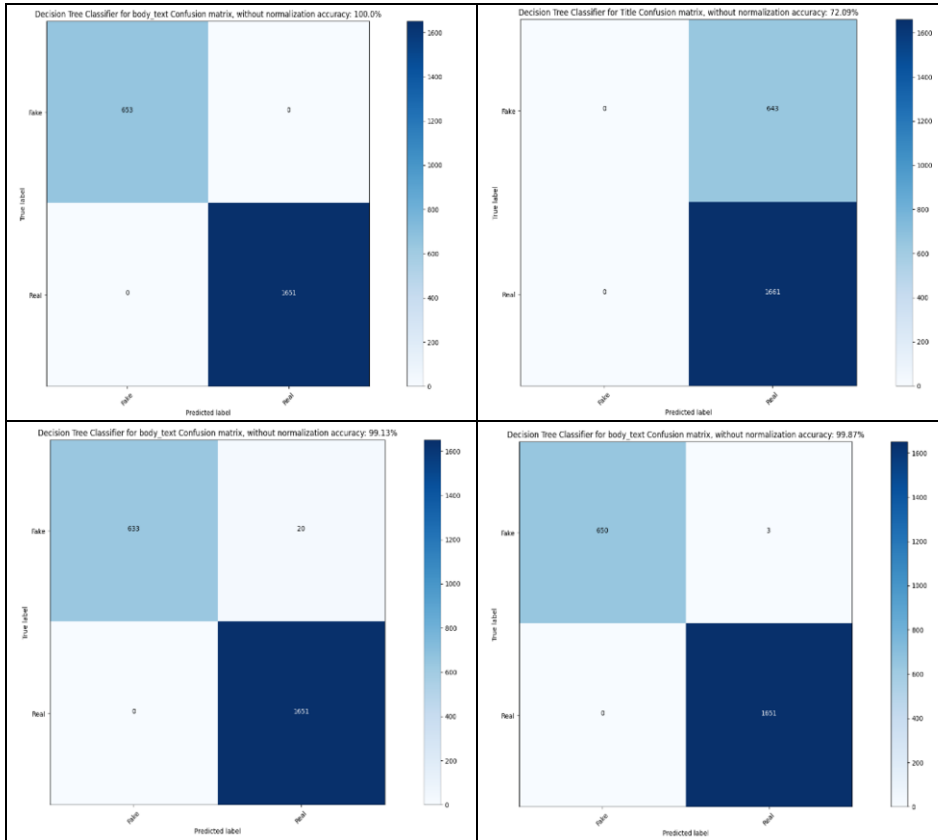

Figure 1. Confusion Matrices for Best and Worst Results

4. Discussion

In this paper, a traditional classifier based on Decision Trees is proposed to assist the Fake News diagnosis from Titles and Full text of an article. The results reveal that the most efficient parameters combination 100% overall accuracy, among the tested is "Entropy" as criterion, "Best: as splitter method, and 40 as max length tree used. However, an interesting point emerges that the application of the algorithm on titles has an overall accuracy of 72,09% on all combination of parameters. This finding reveals that titles are not a reliable part of an article to identify as fake or real. From the algorithmic point of view, this possibly happens since the title has less words comparing to the full body text. From a social point of view, this can lead us to the result that it is not so secure to identify a fake new from the title, but we must read the full article also. The full text containing more information.

5. Conclusions

Nowadays broadcasting media are very powerful on dissemination and proper education to the community on serious global or worldwide risks and dangers, including the appearance of a pandemic case. Broadcasting media can propagate incorrect and antisocial, poor information [17]. Education in the media is not just about Media, but also in any form of expression and communication in the modern, digital era. In related works a variety of the machine learning algorithms applied on tweets from multiple different events to monitor multiple cases of misinformation simultaneously. From the results of this work the model accuracy of various algorithms such as k-Nearest neighbourhood and Support Vector Machines proved efficient for the purpose [18]. In another work a COVID-19 news verification system developed based on three tier system developed (CoVerifi), a frontend based on React.js, a backend based on Python Machine Learning services and a database storing [19]. This work presents an application of Machine Learning tools for the prevention of fake news on health.

References

[1]   Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D, Schudson M, Sloman SA, Sunstein CR, Thorson EA, Watts DJ, Zittrain JL. The science of fake news. Science 2018. Mar 9;359(6380):1094-1096.

[2]   Ireton C, Posetti J. Journalism, Fake News and Disinformation: A Handbook for Journalism Education and Training. France: UNESCO Series on Journalism Education; 2018.

[3]   European Union. Eurobarometer Survey (February 2018): Report-Fake News and Disinformation. Luxemburg: European Union; 2018. ISBN: 978-92-79-81900-1.

[4]   Gómez-González E. Gómez E. Artificial Intelligence in Medicine and Healthcare: applications, availability, and societal impact. Luxembourg: Publications Office of the European Union, EUR 30197 EN; 2020. ISBN 978-92-76-18454-6. doi:10.2760/047666, JRC120214.

[5]   Ganapathy K, Abdul SS, Nursetyo AA. Artificial intelligence in neurosciences: A clinician's perspective. Neurol India 2018;66(4):934-9.

[6]   Raju V, Mohd J, Ibrahim HK, Abid H. Artificial Intelligence (AI) applications for COVID-19 pandemic, Diabetes & Metabolic Syndrome. Clinical Research & Reviews. 2020;14(4):337-339.

[7]   Leroy G, Chen H. Introduction to the special issue on decision support in medicine. Decision Support Systems. 2007;43:1203–1206.

[8]   Wen-Jia K, Ruey-Feng C, Dar-Ren C, Cheng CL. Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. Breast Cancer Research and Treatment 2001;66: 51–57.

[9]   Cristianini N, Shawe-Taylor J. An introduction to support Vector Machines: and other kernel-based learning methods. New York, NY, USA: Cambridge University Press; 2000.

[10]   Haykin S. Neural Networks: A comprehensive foundation. NJ, USA: Prentice Hall; 1999.

[11]   Shakhnarovish D. Nearest-Neighbor Methods in Learning and Vision. USA: The MIT Press; 2005.

[12]   Zhou X, Mulay A, Ferrara E, Zafarani R. ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research. 2020;10.1145/3340531.3412880.

[13]   Scikit-learn. Available at: https://scikit-learn.org/stable/user_guide.html, Accessed 25.1.2021.

[14]   Karlgren J. New Text. Proceedings from the Workshop on New Text: Wikis and Blogs and Other Dynamic Text Sources, held in Conjunction with EACL. ACM, Trento; 2006.

[15]   Bird S, Klein E, Loper E. Natural Language Processing in Python. 2007; Available at: http://nltk.org/.

[16]   Wu H, Luk R, Kwok K. Interpreting TF-IDF term weights as making relevance decisions. ACM Transactions on Information Systems. 2008;26(3).

[17]   Thomas J, Peterson GM, Walker E, Christenson JK, Cowley M, Kosari M, Baby KE, Naunton M. Fake News: Medicines Misinformation by the Media. Clin Pharmacol Ther. 2018 Dec;104(6):1059-1061.

[18]   Hunt K, Agarwal P, Zhuang J. Monitoring Misinformation on Twitter During Crisis Events: A Machine Learning Approach. Risk Anal. 2020 Nov 14. doi: 10.1111/risa.13634. PMID: 33190276.

[19]   Kolluri NL, Murthy D. CoVerifi: A COVID-19 news verification system. Online Soc Netw Media. 2021 Mar;22:100123. doi: 10.1016/j.osnem.2021.100123. Epub 2021 Jan 23. PMID: 33521412.