Paper ###-2020

LightANN: a novel fake news detection approach

Stefano Aparo, Anastasia Marzi, Letizia Orsini, Alessandro Riboni,
University of Milano Bicocca;

ABSTRACT

In the era of social media, sharing information and opinions is becoming more and more easy. The disadvantage of this positive phenomenon is the spread of false information, the so-called "fake news". For this reason, the identification and classification of fake news has become a critical and challenging task. The goal of this paper is to find the best approach to making fake news detention; with this aim, two different frameworks are developed. In the first Classical Fake News Detection (CFND) approach, two different techniques of feature engineering (Bigram and POS-tagging) are applied before feeding the data in three different machine learning algorithms. On the other side, in the "innovative approach", recently developed methodologies are used. In particular, this method is based on a combination of two vectorization techniques (Doc2Vec and document term matrix) and of an Artificial Neural Network (ANN), namely LightANN. The difference between these two approaches is that with the second one is possible to obtain a better result using a limited number of features.
In the CFND approach, the best result in the identification of fake news against a real one is obtained with the method based on Bigram and Linear Support Vector Machine (92% of accuracy). In LightANN, the ANN with document term matrix has the best performance (93% of accuracy).

INTRODUCTION

Nowadays, social media and the Web in general have become one of the most used ways to consume and share information. It is possible to find news online about everything. Social media is cost-free, easy to access, and can fast disseminate posts. Every person that owns a laptop or a smartphone and has access to an internet connection is free to create a post about a particular subject and, sometimes, it is really easy to make others believe the things you are writing. But the problem is exactly that anyone can share their opinion online. And it is here that the spread of false information begins. Since there is no regulatory authority on social media, it also enables the widespread of fake news.
The term "fake news" refers to the false information that is spread deliberately to deceive people (Shu et al., 2018). Fake news as a phenomenon is not new, but in the era of social media, it is becoming easier to distribute information to a global audience. A Google Trends analysis reveals that this term began to gain relevance in US Google searches around the time of the US presidential election in 2016, and remained popular since (Roozenbeek et al., 2019).
It is easy to understand why the detection of fake news is becoming such an important challenge for our society. Anyway, identifying this kind of news is an arduous task. For this reason, the goal of this project is to find the best way to classify news. In order to do this, two approaches are used, one more traditional and one more innovative. The first (CFND) deploys algorithms of Machine Learning, while the second (LightANN) focuses on the implementation of a feedforward Artificial Neural Network. This choice of comparing two approaches is made in order to verify if lowering the level of complexity of the data would still lead to simple and good results. To make the study more realistic, models are developed paying attention to collect homogeneous news in topics and text length but

heterogeneous in sources. Care is taken to create a dataset as close to a real case as possible.

DATA COLLECTION AND DATA MERGE

To make the data more like a real case, the dataset used for the analysis is the result of the merging of three different datasets. All the datasets are retrieved from the platform Kaggle and a certain number of news, both real and fake, are sampled to have a final balanced dataset. This means that at the end of the sampling, the dataset contained almost the same frequence of real and fake news. The datasets used were the following:

- Getting Real about Fake News Dataset, from which 1221 fake news was retrieved;
- All the News Dataset, which was composed by 143,000 real articles from 15 American publications. The sample contained 1221 real news, to balance the first one;
- Fake News Dataset, was a balanced dataset with 20,000 articles, both real and fake. This dataset was used to add more news for the project. In particular, a stratified random sampling is carried out in order to make the word count distributions of the two classes similar in the final dataset. To carry out this procedure and to ensure that the proportions of information brought by the three datasets were not too unbalanced, 5,000 real news and 4530 fake news are sampled.

In the end, the dataset was made of 11992 observations, of which 6221 are real and 5713 are fake.

Title	Text	Label
California Goes One Step Beyond ObamaCare, Proposes Single-Payer Healthcare System to Include Illegal Aliens - Breitbart	California Democrats made a surprise move late Friday to foil President Trump's promise to repeal ObamaCare — by introducing a healthcare system in California...	REAL
Interview: Andy Worthington On Final Push To Close Guantanamo	President Barack Obama has less than 70 days to achieve one of the key goals that will define his legacy: close the Guantanamo Bay military prison...	FAKE

Table 1. Two entries of the final dataset. Three columns: title, text and label.

TEXT PREPROCESSING

After the dataset is complete, the next step is text preprocessing. It is an important step since it transforms text into a more suitable form so that text can be processed by machine learning algorithms in a better way.
Generally, there are 3 main components: Tokenization, Normalization and Noise removal.
The goal of normalization is "to put all text on a level playing fields". In particular, in this phase all characters are converted to lowercase and numbers are removed, as well as extra white spaces. Also, all the punctuation marks are removed from the text.
Another important step to clean the text is the removing of stop words. This means removing from the text all the words that are very common or which have little or no significance. In a text, usually these are the words with the maximum frequency, such as articles or conjunctions.
At last, two very similar text cleaning processes can be applied: lemmatization and stemming. The aim of both of these methods is to obtain the base form of a word. The difference between the two is that the first one converts words to a lemma, which is a real word that can be found in a dictionary, while the second one simply chops the end of the words and the result can be not an actual words. In the analysis, it is decided to apply just the stemming because the results after some trials were better and more useful.
The flowchart in the appendix A shows the pipeline of the text preprocessing.

From the word cloud in figure 1 it is possible to see the results of the application of the preprocessing techniques used. In addition, it can be seen that the most common words are almost the same in the two groups, demonstrating the homogeneity of content between the groups that shows the complexity of the data. This is necessary to make the case study more realistic.

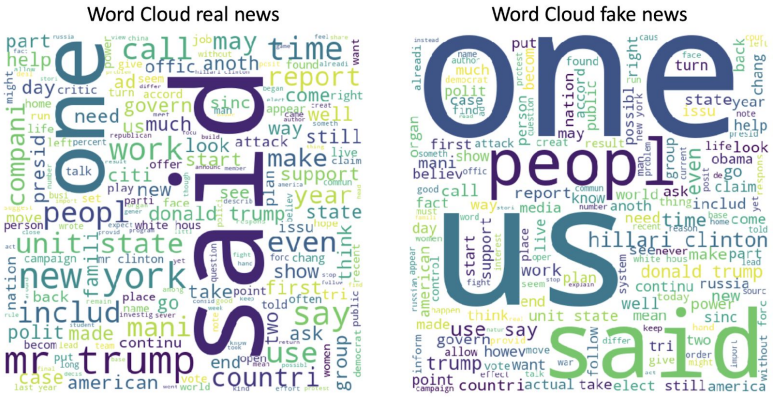

Figure 1. Word clouds of real and fake news after applying text preprocessing.

FEATURES ENGINEERING

One of the most important steps to complete before starting a Machine Learning analysis is the features engineering. It is fundamental to choose the right features in the right format to feed in a model.

After completing the preprocessing phase, the articles are tokenized in individual words. Tokenization consists of splitting a sequence of characters into meaningful units. Then, the words need to be encoded as integers or floating point values for use as input to a machine learning algorithm, called feature extraction (or vectorization).

This phase is carried out with two different Natural Language Processing frameworks. The first one is based on a Classical Fake News Detection (CFND) setting and the second one focus on more innovative method, such as Artificial Neural Network (ANN). The latter, called LightANN, is developed with the aim of decreasing the number of features and maintaining excellent results. See Appendix B for a more detailed description of the frameworks.

In the first technique of CFND approach, the text corpus is converted to a matrix of token counts. Every token is a 2-gram (or bigram). A bigram is a contiguous sequence of two terms from a text. This method has been preferred to unigram because it is able to give the context.

A TF-IDF transformation is applied to the result obtained. This acronym stands for "Term Frequency – Inverse document frequency" and this representation is based on word frequencies. The goal of this transformation is to highlight words that are more interesting. In particular, the importance of the word is proportional to the number of times the word appears in the document but inversely proportional to the number of times the word appears in the corpus (in text mining, the word corpus is used as a synonym of dataset, so the entire collection of text under analysis).

In the second one, feature vectors are created using POS (part-of-speech) tagging applied to single words (unigram), and also in this case a TF-IDF transformation is applied. Part-of-speech tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context.

In the LightANN approach, two methods are developed. Firstly, a Doc2Vec (Le and Mikolov, 2014) encoding is used. Doc2vec is an unsupervised algorithm to generate vectors from sentences or documents. It can be considered an extension of Word2vec that encodes entire documents as opposed to individual words. Doc2Vec vectors represent the theme or overall meaning of a document, and they are word order independent. This technique creates 400 features from the corpus.

Eventually, a further technique based on the use of a document term matrix of the 5000 most frequent words in the dataset is developed. A boolean vector indicates the presence or absence of a word in each document.
Before applying the classification models, the dataset is divided into two parts: the training and the test set. The first one contains 70% of the observations, and the second the remaining 30%.

ANALYSIS

In CFND approach, three machine learning algorithms are applied:

- Logistic Regression is a regression model suitable for the study of binary target variables. So this model is appropriate for the binary classification of real and fake news;
- Random Forest is a heuristic model based on the ensemble of decision trees. It is more potent than simple decision trees because it prevents the risk of oversampling, a typical problem that occurs in text classification;
- Support Vector Machine (SVM) is a model based on separation. In fact, all the models which belong to this group do the classification through a partition of the attribute space. The kernel used is a linear kernel.

With the idea of reducing the dimensionality of the data and still maintaining excellent results, in the LightANN approach, a feedforward artificial neural network is implemented. An Artificial Neural Network is a computational model that is inspired by the networks of biological neurons. A neural network is composed of at least three layers: the first layer is used to draw the linear boundaries, the second allows to combine these boundaries, and the third allows to generate arbitrary complex boundaries. The network uses a method called backpropagation to iteratively update the weights associated with the neurons in the network to minimize the error function. Finally, hyperparameter tuning is done on the main parameters by identifying the optimal configuration to improve the performance. The optimized network is composed of a single hidden layer of 256 neurons. In Appendix C, there is a simplified representation of the architecture of an artificial neural network.

RESULTS

Since the dataset does not present a class imbalance, the accuracy is used as a validation metric to compare the efficiency of the different classifiers.

	Linear SVM	Logistic Regression	Random Forest
Bigram	0.927	0.911	0.863
POS Tagging	0.906	0.885	0.808

Table 2. Model accuracy comparison in CFND approach.

As it is possible to see from Table 2, the classifier that works better in CFND approach is the Linear Support Vector Machine. The model with the highest accuracy is a Bigram with Linear SVM, so this model can correctly understand if a news item is fake or real in 92.7% of cases. For the LightANN approach, the network accuracy using Doc2Vec as feature vectorization technique is 0.913, while using the document term matrix of the 5000 most frequent words is 0.932.
The best methodological frameworks in the two approaches are Bigram with Linear SVM and ANN with Document term matrix. In appendix D, there are the comparisons made through the respective confusion matrices and ROC curves. As one can see from Figure D(i), the SVM identifies slightly better the real news while the ANN gets better results in the fake news classification. Instead, analyzing the ROC curves from Figure D(ii), one can see the performance of the two models is very similar, but the Area Under the Curve of the ANN is greater; this implies a better performance of the model.

GENERALIZATIONS

The project is developed on the comparison of two different approaches. Unlike the CFND, the LightANN allows also capturing the semantic information inside the texts; this means not only to limit the classification according to the presence of words but also to capture the content and meaning of the texts. Moreover, the use of a neural network allows for increasing the performance of the classification model.
The results obtained show that a news item is correctly classified in more than 90% of cases. This can be considered a good result in view of a large scale use to support implementation in fake news detection systems. The tool developed can also be used in real life. For example, it could be added to an application for teenagers. The app may help teens to understand if they should rely or not on what they are reading based on your LightANN approach. Also, it might help journalists that sometimes repost news without knowing the veracity.

FUTURE WORK

One of the possible future implementations is to combine the feature engineering techniques presented by weighing their importance to collect the most significant information from each of them. Moreover, the work done could be applied to data coming from social media (Facebook, Twitter, etc.), also using user information in order to identify profiles that tend to spread fake news.

CONCLUSION

Nowadays, fake news detection is an critical problem. Fake news can influence people's thinking, and this can affect public life and political issues. The two approaches, CFND and LightANN, are applied in the project to try to give a solution to the problem. Both of them lead to an accuracy of over 90%, with LightANN being the best with 93% accuracy.
The LightANN approach limits the number of features: in the first case, only 400 features created through Doc2Vec are used, while in the second one, it uses a document term matrix on the 5000 most frequent words in the text corpus. Finally, the application of an artificial neural network allowed to obtain satisfactory results, since it is obtained on a heterogeneous dataset in the sources and homogeneous in the contents of the two classes, as similar as possible to a real-world scenario.

REFERENCES

Le, Q. and Mikolov, T. 2014. "Distributed Representations of Sentences and Documents" *Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):1188-1196*

Roozenbeek, J. and van der Linden, S. 2018. "The fake news game: actively inoculating against the risk of misinformation*". Journal of Risk Research (2018),* 1–11

Shu, K., Mahudeswaran, D., Wang, S., Lee,D. and Huan Liu. 2018. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *arXiv preprint arXiv:1809.01286 (2018)*.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

- Supervisor: Matteo Borrotti, University of Milano Bicocca, matteo.borrotti@unimib.it
- Stefano Aparo, University of Milano Bicocca,s.aparo@campus.unimib.it
- Anastasia Marzi, University of Milano Bicocca, a.marzi3@campus.unimib.it
- Letizia Orsini, University of Milano Bicocca, l.orsini@campus.unimib.it
- Alessandro Riboni, University of Milano Bicocca, a.riboni2@campus.unimib.it

APPENDIX A: TEXT PREPROCESSING

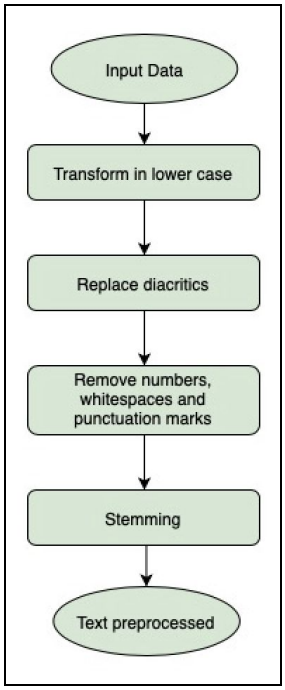

Figure A. Flowchart of text preprocessing

APPENDIX B: FEATURE ENGINEERING

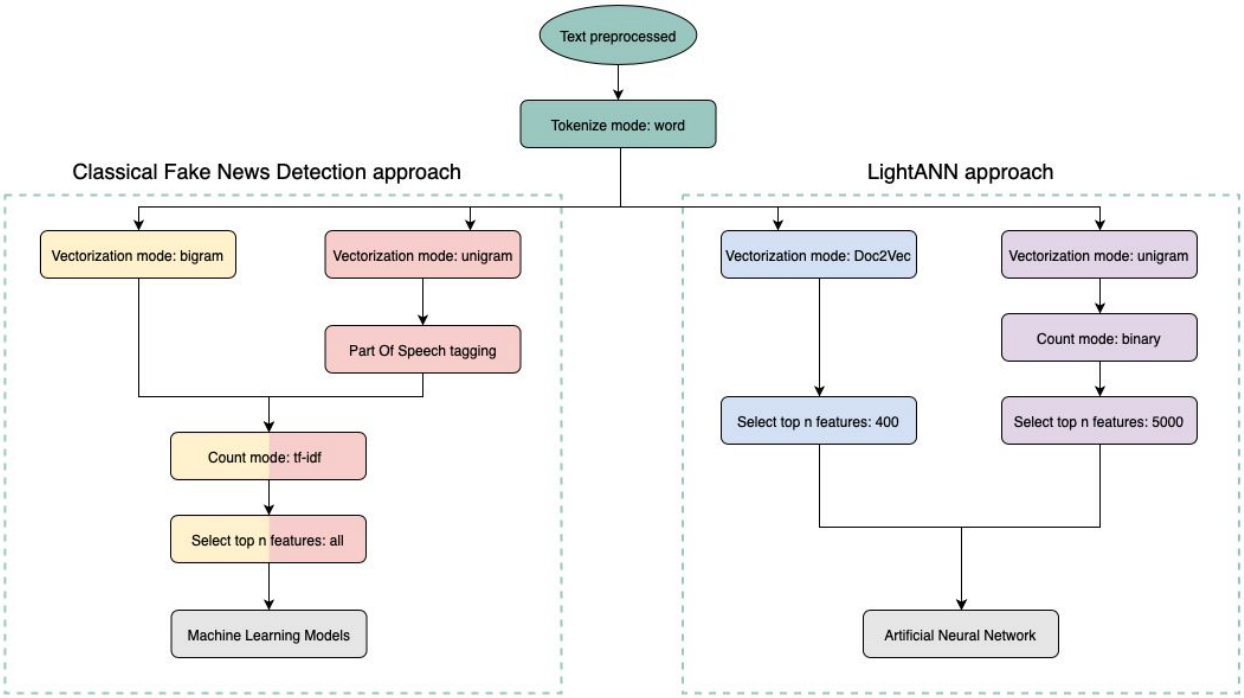

Figure B. Flowchart of feature engineering

APPENDIX C: ANALYSIS

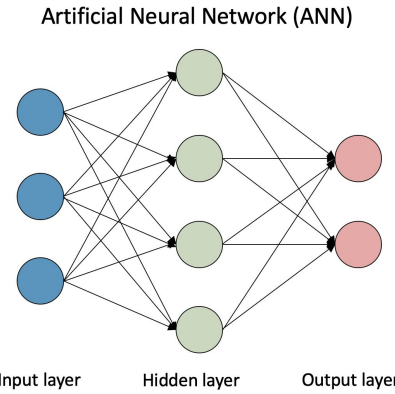

Figure C. Architecture of Artificial Neural Network

APPENDIX D: RESULTS

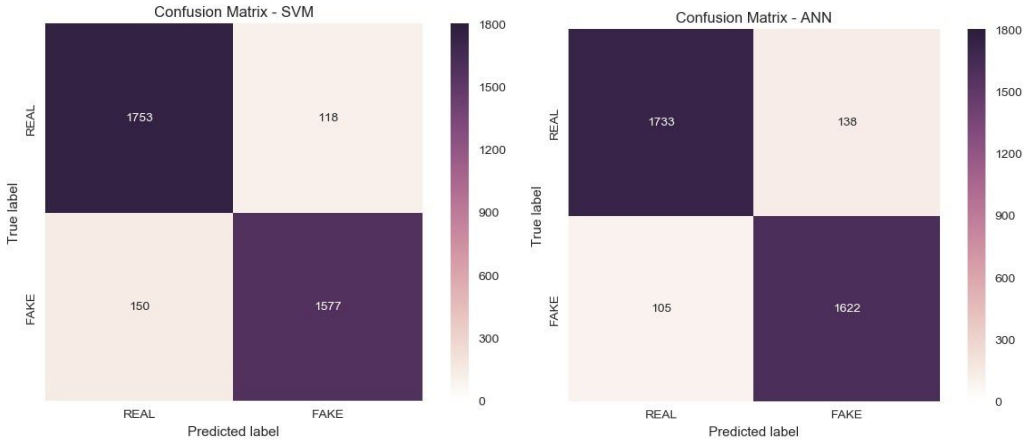

Figure D(i). Confusion matrix of SVM and ANN

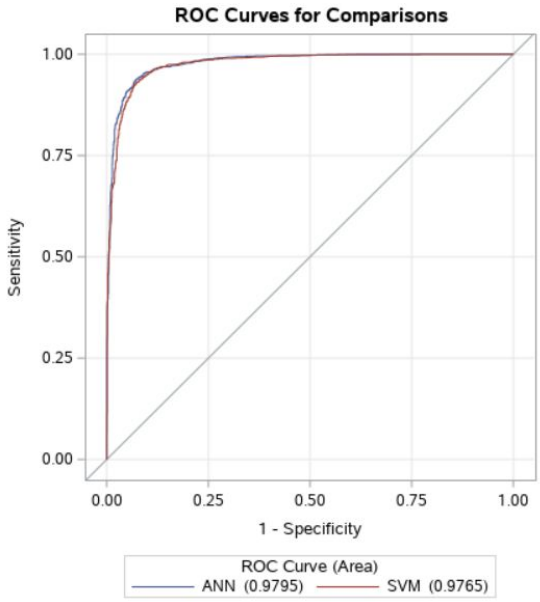

Figure D(ii). ROC curve comparison