PROPOSAL FOR A MODEL FOR DETECTING FAKE NEWS ON SOCIAL MEDIA IN MEXICO

Carlos Augusto Jiménez Zarate\textsuperscript{1} and Leticia Amalia Neira Tovar\textsuperscript{2}

\textit{Universidad Autónoma de Nuevo León, Mexico}

\textsuperscript{1}M. eng.
\textsuperscript{2}Dra

ABSTRACT

The emission of false news on social networks has been increasing continuously, this work analyzes the different automatic detection techniques used for false news, proposes an integration of different machine learning algorithms in addition to the development of a new data set of news tweets in Mexico. To do this, an extraction of tweets from the Mexican media and from sites known as transmitters of false news in Mexico was carried out. The dataset classification test showed that it was the passive-aggressive algorithm that obtained the best accuracy with 79.6%.

KEYWORDS

Fake News, Machine Learning, Social Media

1. INTRODUCTION

The increase in the use of socio-digital platforms has brought with it an increase in harmful practices to social communication, such as the spread of fake news, the use of automated user accounts or profiles known as bots, but there are also the trolls that are users who regularly broadcast provocative and inflammatory content against other users in order to seek self-assertion (Fenoll., 2015).

In the spread of fake news, it is important to be able to analyze the users who are victims of fake news, and some users with certain personality characteristics are more likely to trust fake news (Shu et al., 2019).

Given the magnitude of the spread and spread of fake news, the application of various strategies to try to lessen the impact of fake news has been studied, such as putting warning messages on content categorized as fake news, ratings of the sources influenced the beliefs of the users and established that this rating is a viable measure for fake news (Kim et al., 2017).

But many of these messages have not been shown to have an impact on the detection of fake news by users of social networks (Ross et al., 2019).

There are dimensions such as: Age, gender, education and culture, as factors in the acceptance of fake news, but it is age that has the greatest influence on the acceptance of fake news (Rampersad et al., 2020).

Other users tend to share fake news due to behaviors such as fear of missing something, or fatigue in the use of social networks (Talwar et al., 2019).

The explosive growth of fake news and its impact on different areas such as democracy, justice and public trust has seen an increase in the demand for systems to detect fake news using new datasets, patterns and various latest generation (Zhou et al., 2019).

Perez-Rosas et al (2017) presented two data sets for the detection of fake news that cover seven different news domains, also described the extraction and validation process in great detail, identifying the linguistic differences between fake news and news real.

Kim et al (2018) determined that there is a bias in the information and that most users of social networks cannot distinguish a fake news from a real news, in such a way that, of 83 participants in their research, only one 17% were able to detect a fake news, and only one person was able to detect a fake news more than 60% of the time.
One of the current issues where fake news is having an important impact is in COVID-19, they carried out an investigation, in which they developed a Dataset that includes fake news data from public fact-checking sites such as Politifact, Snopes7 and Boomlive, (Patwa et al., 2020).

Sharma et al (2021) implemented the detection of fake news, using machine learning algorithms, for which they used the “LIAR” dataset from the PolitiFact fact-checking site. For the preprocessing of the data, they started with data cleaning that consists of removing punctuations, eliminating empty words and the stemming function, which allows reducing a word to its root form, for example: Title, Title -> Title. They then implemented function generation to generate a number of characteristics, such as word count, frequency of large and single words, and n-grams. This in order for the algorithms to understand the text and perform the grouping and classification. Within this phase is the vectorization of the data, which is a process of encoding the text as whole numbers. Another vectorization method is the n-gram method quantifies all the combinations of adjacent words or letters of length "n" that can be found in the source text. For the detection of fake news, they used the algorithms of Naive-Bayes, Random Forest, Logistic Regression and finally the Passive-Agressive Classifier. After testing the four algorithms, they determined in their research that the logistic regression algorithm was the one with the highest performance, reaching 75%.

The vast majority of research on fake news has focused on Twitter, but the Naive-Bayes classifier has obtained good results for Facebook, achieving an accuracy of 74% Granik et al (2017).

In addition, there are the artificial positioning strategies known as Astroturfing, which are campaigns that try to influence public opinion through the appearance of viralization of content in socio-digital networks, Astroturfing has two characteristics: The use of deception to hide the origins of the orchestrated campaigns and the lack of transparency of the sponsors of such campaigns (Leiser, 2016).

Khan et al (2019) carried out a benchmark between different pre-trained advanced linguistic models of Machine Learning for the detection of fake news, such as:

- The BERT (Bidirectional Encoder Representations from Transformers), which is a model for learning contextual representations of unlabeled words. They focused on BERT-Base which has 12 layers with 12 attention heads and 110 million parameters.
- RoBERTa (Optimized Focus). This model achieves better performance by using mini batches to train the model for a longer time through more data.
- DistilBERT. It is a smaller, faster and cheaper and lighter version of the original BERT, they have 40% less parameters than the BERT-Base.
- ELECTRA (Efficiently LEarning an Encoder that Classifs Token Replacements Accurately). This self-supervised learning model of language representation. This model takes an input text and randomly masks the text with an input token, ELECTRA is trained to distinguish real input tokens from fake input tokens.
- ELMO (Language model embeddings). It is a contextualized word representation of a deep bidirectional model that is trained on a large text corpus. It has 2 layers and 93.6 million parameters.
- Zervopoulos et al (2020) used various Machine Learning techniques, such as Naive Bayes, Suppport Vector Machine, C4.5 and Ramdon Forest, to be able to classify the linguistic characteristics of fake news, for this they took the tweets in English and Chinese from a Twitter database to classify fake news.
- Gao et al. (2016) evaluated the performance of the convolutional neural network (RNN) algorithm and for the detection of rumors they determined a performance of 0.827.
- Oshikawa et al (2018) implemented an investigation where they used various models of natural language processing (NLP) for the detection of fake news, others such as Cueva et al (2020) compared the natural language processing technique with various models of Artificial Intelligence, among them the Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU) for this they built a Dataset that was based on a data set from the Kaggle platform, which offers various data repositories for different types of projects, including the detection of fake news.
- Liu et al (2020) (2020) presented research for the early detection of fake news, through a neural network model which they named “Fake News Early Detection” or FNED, which is composed of three components. A multiple response function extractor that takes into account the text of responses and the user's profile, a mechanism that highlights important answers, and a grouping mechanism to perform feature aggregation.
Ksieniewicz et al (2019) exposed the different types of digital analysis that can be applied to a news item to detect whether or not it is fake news, among them, the reputation of the author of the news, an image analysis, to determine the context and the detection of manipulation, the analysis of text using natural language processing techniques, psycholinguistic or non-linguistic analysis, and finally the analysis of the news metadata. In addition, they emphasized the impact of fake news during the current COVID-19 pandemic, such is the case of fake news that claimed that “5G” cell phone antennas were causing the coronavirus.

Reis et al (2019), worked to determine new functions for the detection of fake news, also addressed the different analysis topics that can be performed on the text, such as:

- Language characteristics (syntax) and they analyzed the characteristics with methods such as the bag of words, n-grams, labeling, number of words and syllables per sentence.

Monti et al (2019) developed a model for the detection of fake news focused on the study of its spread, taking into account data such as content, user profile and user activity. They determined that the structure and propagation of social networks are relevant characteristics that allow the adequate detection of fake news, and they also established that fake news can be detected efficiently in the first stages of propagation or diffusion.

2. DEVELOPMENT AND RESULTS

Figure 1 shows the proposed model to classify news as false or true, where the text of the tweet and the users who broadcast the content (news) are classified using machine learning algorithms. To find out if the content of a news item is false or real. To determine if a tweet is viral or not, social network analysis will be used, through which it will be determined by statistics of the number of likes by adding the number of retweets obtained for each tweet.

![Fake news detection model in social media in Mexico](image)

Figure 1. Fake news detection model in social media in Mexico

For this paper, only the first part will be addressed, which is the classification of the text of the tweet into fake news or real news, for which the construction of a dataset of news content broadcast on Twitter Mexico was carried out.

For the development of a dataset of Mexican media tweets, various Mexican news media accounts that have a Twitter account were taken into account; the Twitter API was used to extract the tweets.

To extract tweets, it is necessary to have the api keys provided by Twitter and establish the keywords or accounts from which the streaming will be carried out. Figure 2 shows the process for the development of the data set related to the news media in Mexico that have a Twitter account.
Once the tweets were obtained, they were classified manually, for which a team of volunteers was formed, who were shown the definition of the Cambridge dictionary that says that fake news are: “False stories that seem to be news, spread on the internet or using other means, generally created to influence political opinions or as a joke”. For the definition of news, the dictionary of the Royal Spanish Academy was used, which says the news is: Information about something that is considered interesting to disseminate; disclosed fact or new information regarding a matter or a person.

The total number of tweets extracted was 5,108 tweets, of which 1,126 tweets were manually classified in the two categories proposed for this work that are “fake news” or “real news”.

This data set will be used to form the training data set. Table 1 shows us the classification percentages of the training data set, it can be seen that the percentage of tweets classified as false news represents 52.8% and the percentage of tweets classified as real news represents 47.1%, thus obtaining a balanced dataset.

Table 1. Types of tweets in the dataset (manual classification)

<table>
<thead>
<tr>
<th>Tweets type</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fake news</td>
<td>52.8</td>
</tr>
<tr>
<td>Real news</td>
<td>47.1</td>
</tr>
</tbody>
</table>

Table 2 shows the structure of the dataset and the text of some of the tweets that were classified for the development of the new dataset of news tweets in social media in Mexico. It can be seen that it is composed of two columns, the first refers to the body of the tweet or plain text, and the second column is the label with which the content of the tweet or text was categorized.

Table 2. Sample of news tweets that were classified as false news (Fake) or real news (True)

<table>
<thead>
<tr>
<th>Tweets text</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peña Nieto sabía que Juan Gabriel estaba vivo pero le pidió ocultarse ha... <a href="https://t.co/bHK2IW9YRk">https://t.co/bHK2IW9YRk</a> vía @YouTube</td>
<td>FAKE</td>
</tr>
<tr>
<td>Captaron a Juan Gabriel en la marcha del orgullo gay 2019! <a href="https://t.co/drEe70MqdF">https://t.co/drEe70MqdF</a> vía @YouTube</td>
<td>FAKE</td>
</tr>
<tr>
<td>Solo llegaron tres mil dosis. El gobierno y Morena no paran de hacer el ridículo. <a href="https://t.co/hzLr0eUNUR">https://t.co/hzLr0eUNUR</a></td>
<td>FAKE</td>
</tr>
<tr>
<td>AMLO autoriza nuevo partido de Elba Esther y le da $20 mil millones de ... <a href="https://t.co/zE9N0hfrso">https://t.co/zE9N0hfrso</a> vía @YouTube</td>
<td>FAKE</td>
</tr>
<tr>
<td>El primer mandatario indicó que este “es el país con más #fraudes electorales en la historia... se tienen que terminar... <a href="https://t.co/BAWcgXO5r7">https://t.co/BAWcgXO5r7</a></td>
<td>TRUE</td>
</tr>
<tr>
<td>El fútbol dará inicio a los Juegos Olímpicos de Tokio el miércoles, con el enfrentamiento entre Japón y Australia... <a href="https://t.co/YS4ENYzYov">https://t.co/YS4ENYzYov</a></td>
<td>TRUE</td>
</tr>
<tr>
<td>Funcionarios de la @Registraduria de #Colombia reconocieron el trabajo que realiza el @iecm en materia electoral... <a href="https://t.co/upP014m6O0X">https://t.co/upP014m6O0X</a></td>
<td>TRUE</td>
</tr>
<tr>
<td>Guerrero retrocedió de semáforo verde a amarillo ante repunte de casos de #Covid19, informó el Gobernador Héctor As... <a href="https://t.co/0PqQRlqjsxk">https://t.co/0PqQRlqjsxk</a></td>
<td>TRUE</td>
</tr>
</tbody>
</table>
There are various classification techniques such as:

The Naive-Bayes algorithm (NB). This algorithm is a posterior probability classifier. This algorithm assumes that for a sample $X$, its attributes $X_1, X_2, ..., X_n$ have a linear independence with respect to the value of the class, so that the conditional probability is presented as the product of the conditional probabilities of the attributes separately. This algorithm assumes that the occurrence or non-occurrence of any characteristic is not related to the occurrence or non-occurrence of any other characteristic (Frank et al., 2000).

Logistic regression (LR). This algorithm works in a similar way to linear regression, but they have a binomial response variable, and it is also capable of modeling probability based on individual characteristics. But since the probability is a ratio, what is modeled in this algorithm is the logarithm of the probability (Sperandei, 2014).

Multi-layer Perceptron (MLP). It is a supervised learning algorithm that learns from a function training on a dataset where "m" are the input dimensions and "o" are the output dimensions. Unlike logistic regression between the input and output layer there can be one or more non-linear layers called hidden layers. Although this algorithm is very common, it is also necessary to take into account some of the problems, such as the choice of parameters and network configuration, if there is a small learning rate it can lead to a slow convergence, but if there is a slow learning rate high this can cause jumps in the solution (Popescu, 2009).

Support Vector Machine (SVM). This algorithm can be described as a surface learning algorithm of two distinct classes of entry points. The SVM tries to draw a hyperplane to be able to classify vectors with a high degree of similarity (Evgeniou & Pontil, 2001).

Passive-aggressive (PA). This algorithm is also known as passive-aggressive online; it has a large margin for online ranking, they are called passive because if the prediction is correct, it maintains the representation and does not make any exchange and aggressive because if the prediction is incorrect, it makes changes in the representation, that is to say, some exchange in the representation can correct the prediction; its main parameters are: C. It is to formalize and denotes the penalty that the model will make in a prediction error; max_iter. It is the maximum number of duplications that the model performs on the training data (Suganthi, 2021).

Table 3 shows the precision results of the algorithms proposed for the fake news detection model, being the passive aggressive algorithm (PA) the one that obtained the best result for the test dataset of news tweets in Mexico, with a 79.6%, followed by the Naive Bayes (NB) algorithm who obtained an accuracy of 78.2%, the algorithm that obtained the lowest score was the multi-layer perceptron (MLP) with 71.6%. For this paper the user analysis and the analysis of the propagation of the tweets are left for a future delivery. However, it is proposed to perform the analysis of the user profile through the Botometer interface tool which uses a machine learning model, it also offers an API to integrate the analysis of several accounts through a simple open code and free access.

Table 3. Accuracy of algorithms for the classification of tweets of the proposed model for the detection of fake news on social media in Mexico

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive-Bayes (NB)</td>
<td>78.2%</td>
</tr>
<tr>
<td>Logistic regression (LR)</td>
<td>75.6%</td>
</tr>
<tr>
<td>Multi-layer Perceptron (MLP)</td>
<td>71.6%</td>
</tr>
<tr>
<td>Support vector machine (SVM)</td>
<td>76.0%</td>
</tr>
<tr>
<td>Passive-aggressive (PA)</td>
<td>79.6%</td>
</tr>
</tbody>
</table>
3. CONCLUSION AND FUTURE WORK

The model presented proposes the integration of different algorithms of machine learning and integrating it with other analyzes such as the user profile and the level of spread or viralization of the news tweets. For this, a totally new data set was developed with tweets from the year 2021, made up of 1,126 tweets that were manually classified into real news (True) and false news (Fake). The test of the data set showed that it was the passive-aggressive algorithm that obtained the best classification result with 79.6% accuracy. Although the results are below 95%, it shows that the development of a new dataset of news tweets in Mexico is feasible. As future work, work will be done on the expansion of classified content, and the manual classification process will be reviewed to increase accuracy, the integration of the analysis of user profiles and the level of propagation or viralization; and thus, be able to implement the false news detection model in social networks in Mexico.

REFERENCES


