

Detecting fake news using Twitter social information

Jesús M. Fraile-Hernández*, Álvaro Rodrigo and Roberto Centeno

NLP & IR Group at UNED (Spain)

Abstract

In this paper, the aim is to study whether social information can provide useful information when classifying news. For this purpose, a set of news items in Spanish has been extended with social information. Subsequently, a classifier model has been proposed to carry out this task, mixing the social information previously extracted with the textual information of the news item. Finally, we have studied which social features are the most relevant in this task.

Keywords

Social information, Classifying news, Classifier model, Social features, Fake news detection

1. Introduction

Due to the increase in communication channels in recent decades, users have access to an immense amount of information almost instantaneously. However, it is relatively easy to fall for hoaxes or misinformation on social media.

Traditional models of fake news detection focus on detecting the linguistic characteristics of the news. Subsequently, in [1], pre-trained embeddings were used along with LSTM. Finally, with the emergence of contextual models, [2] leveraged the pre-trained BERT model, to perform transferred learning and identify the veracity of news.

However, due to the difficulty even for a human to discern between true and false news, sometimes the textual information in the news is not enough. In [3] it is proposed at a theoretical level the possibility of creating a hybrid approach that incorporates the linguistic characteristics of the news and an analysis of the networks that are formed around that news. In [4] the author uses different features to identify fake news in popular Twitter threads. In [5] fake news is detected using only the extracted textual information. Regarding hybrid models, the CSI model proposed in [6] performs a characterisation in three modules: capturing, scoring and integrating. In [7], a news detection model is proposed that considers the association of user interactions, the editor's bias and the users' stance towards the news.

The aim of this work is to study whether social information can provide useful information for the detection of fake news. To this end, social information has been collected from Twitter to extend FakeDeS, a relevant corpus of news in Spanish, and a model has been designed to include textual and social information. Furthermore, we intend to study which social features are the most relevant for news classification.

The rest of this paper is structured as follows: Section 2 describes the datasets to be used along with the task to be solved. Section 3 describes the methodology followed including the extraction of social information from Twitter along with the models proposed based on the data they use. Section 4 includes the evaluation metrics used. Section 5 then presents the results, which will be discussed in Section 6. Finally, conclusions and future work are given in Section 7.

Proceedings of the 1st Workshop on COuntering Disinformation with Artificial Intelligence (CODAI), co-located with the 27th European Conference on Artificial Intelligence (ECAI), pages 19–28, October 20, 2024, Santiago de Compostela, Spain

*Corresponding author.

✉ jfraile@lsi.uned.es (J. M. Fraile-Hernández)

ORCID 0009-0001-5474-4844 (J. M. Fraile-Hernández)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

| Team | Fake | True | F_{macro} | Accuracy |
|------------------------------------|---------------|---------------|---------------|---------------|
| GDUFS_DM | 0.7666 | 0.7649 | 0.7666 | 0.7657 |
| Haha | 0.7548 | 0.7522 | 0.7548 | 0.7535 |
| Chats_ | 0.7514 | 0.7690 | 0.7514 | 0.7605 |
| SINAI | 0.7385 | 0.7821 | 0.7385 | 0.7622 |
| baseline-BERT | 0.7321 | 0.7432 | 0.7321 | 0.7378 |
| baseline-BOW-SVM | 0.7217 | 0.7359 | 0.7217 | 0.729 |
| Lcad_UFES | 0.7102 | 0.6837 | 0.7102 | 0.6976 |
| CITIUS-NLP | 0.7098 | 0.4940 | 0.7098 | 0.6311 |
| baseline-CHAR-3-GRAMS-SVM | 0.7063 | 0.6883 | 0.7063 | 0.6976 |
| zk15120170770 | 0.7053 | 0.6053 | 0.7053 | 0.6626 |
| ForceNLP | 0.6925 | 0.4739 | 0.6925 | 0.6119 |
| GRX | 0.6915 | 0.5624 | 0.6915 | 0.6381 |
| TSIA | 0.6860 | 0.5263 | 0.686 | 0.6224 |
| FREE | 0.6855 | 0.6519 | 0.6855 | 0.6696 |
| LIMCA | 0.6812 | 0.7027 | 0.6812 | 0.6923 |
| ZZWEI | 0.6737 | 0.6794 | 0.6737 | 0.6766 |
| Premjithb | 0.6576 | 0.7177 | 0.6576 | 0.6906 |
| Sdamian | 0.6542 | 0.75 | 0.6542 | 0.7098 |
| Yeti | 0.6316 | 0.609 | 0.6316 | 0.6206 |
| Gulu | 0.6226 | 0.476 | 0.6226 | 0.5612 |
| Nicksss | 0.6119 | 0.7592 | 0.6119 | 0.7028 |
| Bribones tras la esmeralda perdida | 0.5835 | 0.5878 | 0.5835 | 0.5857 |
| WSSC | 0.5118 | 0.6657 | 0.5118 | 0.6031 |
| Skblaz | 0.4838 | 0.649 | 0.4838 | 0.5822 |

Figure 1: Results IberLEF 2021 on the test set.

2. Dataset and task

The dataset we will work with is the Spanish Fake News Corpus (FakeDeS) [8], which contains publications in Spanish about different events that were collected from November 2020 to March 2021. Each of these publications is labelled as true or false. Newspaper websites and fact-checking websites were mainly used to collect the information.

The dataset is divided into 3 files with a total of 1543 news items. Because of the methodology used, it has been decided to merge the training and development files to obtain what we will call the training set. Each of the news items contains information such as the topic, the name of the source, the headline, the text and the link to the news item.

The training set has a total of 971 news items, of which 480 are false and 491 are true. On the other hand, the test set consists of 572 news items, half of which are true and half of which are false. Therefore, we are dealing with balanced data sets.

The topics covered in the training corpus are: politics, entertainment, sport, society, science, health, economy, security and education.

It should be noted that the test set has news related to Covid-19, while the training set does not present any news related to this topic (the most similar are the health news, but in no case do they mention Covid-19). Therefore, the models that are proposed will have to correctly classify this topic without having seen it in the training.

In IberLEF 2021, a shared task was proposed whose objective was to classify a series of news items as true or false. To do so, the FakeDeS corpus described above was used. A report was published in [9], which collected the most important characteristics of the best-performing models. The results of this task by the different participants can be seen in Figure 1. Among the approaches used to solve it, the participants of the GDUFS team, the team that achieved the best accuracy, used a BERT model and sample memory with an attention mechanism. The method consisted of taking the first and last segments of the texts and feeding them into a BERT system, obtaining two embeddings (head and tail). In addition, there is a matrix called ‘sample memory’, which is obtained by taking a random sample of the head and tail embeddings; this matrix is used in an attention mechanism with the rest of the

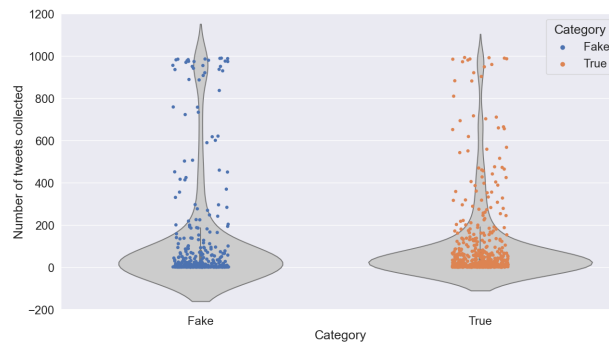


Figure 2: Violin diagram of the number of tweets collected.

texts. In contrast to the GDUFS_DM approach, the participants of team Haha, the second-placed team, employed feature selection with a weighted tf-idf and a multilayer perceptron. This model not only analysed the content of the news item, but also combined information such as the publisher of the news item or the topic of the news item.

3. Methodology

This section describes the methodology used to extract social information from Twitter users. In addition, the models trained according to the type of data they use are presented.

3.1. Social information extraction

The main objective of this work is to study the information provided by social information when detecting fake news, and as mentioned in Chapter 1, there is no corpus in Spanish that contains this information. This is why we decided to extract this information from the social network Twitter, using the API provided by the platform.

For each news item, we searched for those tweets that contained the headline of the news item or the link to it. To solve the problem of the maximum length of the queries, special characters have been eliminated from the news headlines.

According to [4] and [5] there is a series of metadata of the tweets that allow extracting information about whether the user may be prone to the propagation of fake news or the tweet may contain untruthful information. Therefore, it has been decided to extract the following metadata from each of the tweets.

- **Tweet.** Text of the tweet, id of the author, id of the tweet, number of retweets, number of replies to the tweet, number of likes, number of citations of the tweet.
- **User.** username (*str*), user creation date (*date*) *ISO 8601*, verified user (*bool*), number of followers (*int*), number of followed (*int*), number of tweets (*int*), number of times listed (*int*).

We have managed to extract posts from 41.67% of the total number of news items. Of these, the distribution of the number of tweets collected per news item shows a high concentration in the (0, 200) interval, representing 86% of the news items. Within this interval, it is observed that true news tends to receive more interaction. However, as the number of tweets about a news item increases, it is evident that fake news receives a greater number of interactions. This trend can be seen in the violin diagram presented in Figure 2.

It is worth noting that, although the news is written in Spanish, there are tweets in English or French that talk about the news. This is especially true for news related to Covid-19.

3.2. Textual models

In this section, the textual methods used for the binary classification of the news items will be presented. The full text of the news item has been used, so it has had to be preprocessed. For the non-contextual models, urls, emoticons or non-textual expressions, stopwords, the text has been converted to lowercase and the processes of lemmatisation and stemming have been applied. However, for the contextual models, only the urls have been eliminated.

Subsequently, 5 different approaches have been used.

1. Vector space model based on bags of words (BoW).
2. Vector space model using a weighted tf-idf.
3. Bigram counting.
4. Neural Networks and deep learning.
5. Contextual models.

For approaches 1, 2 and 3, Naive Bayes, SVM, Logistic Regression, Decision Trees and Random Forest models have been trained. For approach 4, multilayer perceptrons with input the tf-idf weight vector, multilayer perceptrons and convolutional networks with an embedding layer and multilayer perceptrons, convolutional networks, LSTM, GRU and bidirectional networks with a pre-trained embedding layer. Finally, for approach 5, the BETO model has been selected: Spanish BERT [10] with a final classification layer with two neurons. This model is a BERT model trained with the whole-word masking technique on a large corpus of more than three billion Spanish words.

3.3. Models with social information

The methods that use only the social information of the news collected use the following metadata for each published tweet: number of retweets, number of replies, number of likes of the tweet, number of quotes of the tweet, verified user, number of followers, number of followed, number of tweets of the author, number of times the author has been listed. Then, in order to record the impact of the news item on social networks, the number of tweets collected for this news item is added.

To represent all the tweets that talk about a certain news item, an average of the previous characteristics of each tweet has been calculated. Finally, the standard deviation of each characteristic was added. In this way, a data matrix with 20 columns is obtained (where the column relating to the deviation of the number of tweets of the news item is always 0).

Once the feature matrix has been obtained, different learning models have been used with different hyperparameter explorations such as Decision Trees, Random Forest, SVM, Gradient Boosting, Adaptive Boosting, MLP,...

3.4. Hybrid model

A hybrid model has been developed that seeks to take advantage of both the textual information provided by the text of the news item and the social information extracted from the Twitter data (both the non-textual information of the section and the text of the tweets collected).

In this model, for each news item, a specialised model is used to classify the news using social information. For this purpose, the best model from the previous subsection (Random Forest) is selected. With this model, for each news item, the probabilities of being true or false are extracted using as input the corresponding row of the matrix of social characteristics with standard deviation described in that section. In the event that no tweets could be extracted from a news item, the output would be a vector of two zeros.

In parallel, the text of the news item is processed using the BETO: Spanish BERT model [11]. The output is a vector of dimension 768.

In parallel to these two processes, for each news item with tweets collected, the text of each tweet is pre-processed (eliminating URLs and tokenising) and subsequently processed using the pre-trained XLM-roBERTa-base model [12]. This transformer model has been trained on a corpus of about 198

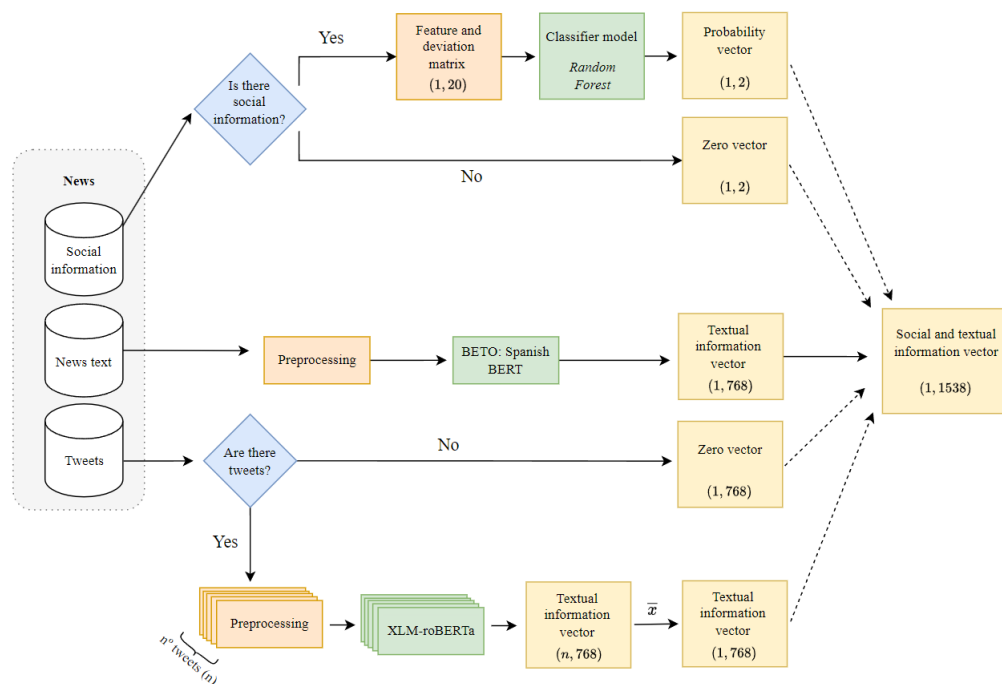


Figure 3: Workflow of the hybrid model.

million tweets in 8 different languages (Spanish, Arabic, English, French, German, Hindi, Portuguese and Italian) and is specialised in sentiment classification (positive, negative or neutral). In our case, the last layer of the model will be removed, obtaining as output a vector of length 768 that will represent the most relevant features of the text of the tweet.

For each available tweet, the previous process has been carried out, obtaining a vector of length 768. Finally, an average of all the vectors of the tweets of the news item has been made to obtain a vector that represents the tweets of that news item. If the news item had no social information, a vector of zeros is returned.

Then, the three vectors are joined to obtain a vector of dimensionality 1538. This flowchart can be seen in Figure 3

Once all the news has been processed following the previous diagram, several models have been trained such as Decision Trees, Random Forest, SVM, Gradient Boosting, Adaptive Boosting, MLP, ...

4. Evaluation

Two different methodologies have been used to evaluate the models, a cross-validation and an evaluation on the test set.

4.1. k -fold cross-validation

Cross-validation is one of the most widely used methods to estimate the prediction error of a model with a given set of hyperparameters. A k -fold (or k -fold cross-validation) has been used. This method divides the data set, in our case the train set together with the development set, into k equal parts P_1, \dots, P_k . For each P_n the model is trained using the other $k - 1$ parts and the error in predicting the P_n data (data never seen by this model) is calculated. By doing this for the k parts we obtain a set of errors. With these k errors we calculate their mean and variance to obtain a measure of the average error of that model with those hyperparameters.

It should be noted that this method requires a fairly large computational cost, since for a cross-validation of k -folds it would be necessary to train k models. As a general rule, a value of 5 or 10 is

| Textual models | F_1 |
|-----------------|--------------|
| TF-IDF (RF) | 0.849 |
| BoW (RF) | 0.825 |
| Bigramas (RF) | 0.822 |
| MLP (Embedding) | 0.786 |
| MLP (TF-IDF) | 0.751 |
| CNN | 0.740 |
| BETO | 0.727 |
| GRU | 0.678 |

Table 1

Cross-validation results of textual model training.

usually chosen as a good compromise between bias and variance. In our case a 5-fold cross-validation has been used.

4.2. Test set evaluation

Finally, for the model that has performed best in the previous cross-validations, the test set will be evaluated. This set will never be seen by the model and will provide a representation of the generalisability of the model.

4.3. Evaluation metrics

To evaluate the performance of our classification model, we use the F1 metric. The F1 value will be calculated for both true and false classified news. With this, the value *Macro-F1*, or simply *F1*, will be calculated as the average between the two previous values.

5. Results

In this section the results of the various trained models will be presented. For each approach in the section 3 the following results will be shown:

- Within the training of a particular approach, the *Macro-F1* value of the best algorithms used will be shown. The average of the *Macro-F1* values will be reflected using 5-fold cross-validation.
- For each approach, the model with the best *Macro-F1* will be selected during training. Subsequently, it will be retrained with all data and evaluated on the test set. The $F1_{Fake}$, $F1_{True}$, *Macro-F1* and the Accuracy of the model will be exposed.

5.1. Textual models

The training results of the methods described in section 3.2 are listed in Table 1.

It can be seen that the non-neural models stand out from those using neural networks. This could be due to the fact that the models being used have a large number of parameters to optimise and we have a rather limited data set. It is worth noting that the use of pre-trained embeddings has resulted in lower performance than training the embeddings from scratch. Also noteworthy is the poor performance obtained with recurrent networks, models that have required a large amount of training time and are commonly used for language processing problems. The best performing approach has been to use a weighted tf-idf together with a Random Forest model.

The results of the evaluation of this model on the test set and the results of the teams participating in IberLEF 2021 are shown in Table 4.

| Social information models | <i>F1</i> |
|------------------------------------|--------------|
| Random Forest | 0.845 |
| Gradient Boosting | 0.834 |
| Adaptive Boosting | 0.826 |
| Extremely Randomized Trees | 0.817 |
| Decision Trees | 0.797 |
| K-Nearest | 0.788 |
| Multilayer Perceptron (MLP) | 0.787 |
| SVM | 0.785 |
| Passive-Aggressive Classifier | 0.785 |
| Perceptron with two hidden layers | 0.783 |
| Linear Discriminant Analysis (LDA) | 0.781 |
| Multinomial Naive Bayes | 0.781 |
| Perceptron with one hidden layer | 0.781 |
| Bernoulli Naive Bayes | 0.779 |
| Quadratic Discriminant Analysis | 0.776 |
| Logistic Regression | 0.703 |

Table 2

Cross-validation results of social models.

5.2. Social information models

The training results of the methods described in section 3.3 are collected in Table 2.

We can see that the *F1* of the models is quite high. Tree-based models occupy the top 5 positions in the list. In addition, those based on clusters of trees stand out from individual decision trees. The best performing approach was a Random Forest model. It should be remembered that this model has only been trained and evaluated with those news items from which it has been possible to extract social information, so the training and test set is smaller than in the rest of the cases.

Due to these results, it has been decided to choose the Random Forest classifier for the social information for the hybrid model, as indicated in section 3.4.

5.3. Hybrid model

The training results of the methods described in section 3.4 are listed in Table 3.

In view of the training results, any of the first 2 models would be valid for your choice. The rest of the models have a very similar accuracy to the first three. It has been decided to select logistic regression over decision trees since it is a simpler algorithm, with a smaller number of hyperparameters and with a lower computational cost.

The results of the evaluation of this model on the test set and the results of the teams participating in IberLEF 2021 are shown in Table 4.

6. Discussion

This section presents a discussion of the results obtained.

In view of the results shown in Tables 1 and 3, it can be seen that the approach that obtains the best *F1* is a model that uses only textual information, more specifically a Random Forest with a weighted tf-idf. This approach obtains a higher *F1* compared to other types of models that include social information, so that a priori it could be thought that social information does not provide relevant information.

However, in Table 4, we can see how on the test set the model that uses only textual information obtains worse results compared to the hybrid model. This is due to the fact that when using a tf-idf weight it is possible that there are words in the corpus on which the weight is applied (training news corpus) that do not exist in the test set. This is why models such as transformer networks pre-trained

| Hybrid Model | F1 |
|-------------------------------------|--------------|
| Decision Trees | 0.818 |
| Logistic Regression | 0.818 |
| SVM | 0.809 |
| Linear Discriminant Analysis | 0.809 |
| Random Forest | 0.809 |
| Gradient Boosting | 0.809 |
| Passive-Aggressive Classifier | 0.809 |
| Adaptive Boosting (AdaBoost) | 0.809 |
| Extremely Randomized Trees | 0.809 |
| Quadratic Discriminant Analysis | 0.809 |
| Multilayer Perceptron (MLP) | 0.809 |
| K-Nearest | 0.809 |
| Perceptron with three hidden layers | 0.809 |
| Perceptron with two hidden layers | 0.808 |
| Perceptron with one hidden layer | 0.808 |
| Multinomial Naive Bayes | 0.631 |
| Bernoulli Naive Bayes | 0.607 |

Table 3

Cross-validation results of hybrid model.

| | Fake | True | F_{macro} | Accuracy |
|------------------|---------------|---------------|---------------|---------------|
| Textual Models | 0.7140 | 0.7488 | 0.7314 | 0.7325 |
| Hybrid Model | 0.7900 | 0.7352 | 0.7626 | 0.7657 |
| GDUFS_DM | 0.7666 | 0.7649 | 0.7666 | 0.7657 |
| Haha | 0.7548 | 0.7522 | 0.7548 | 0.7535 |
| Chats_ | 0.7514 | 0.7690 | 0.7514 | 0.7605 |
| SINAI | 0.7385 | 0.7821 | 0.7385 | 0.7622 |
| baseline-BERT | 0.7321 | 0.7432 | 0.7321 | 0.7378 |
| baseline-BOW-SVM | 0.7217 | 0.7359 | 0.7217 | 0.7290 |

Table 4

Results on the test set. Including the best participants of IberLEF 2021.

on large corpora will have more generalisation capacity and, therefore, will be able to obtain better results.

Once social information is introduced into the model, a significant increase in results can be seen. This is due to the fact that on the one hand the text is being processed using transformer models with a very high generalisation capacity and that the non-textual social information extracted from Twitter is the same regardless of the subject matter.

Comparing the models with respect to the best classified in IberLEF 2021, Figure 1, it can be seen that the hybrid model is the one that best classifies Fake news. This hybrid model obtains the same Accuracy as the first ranked team.

In addition, a study has been carried out on which social information features are the most relevant for the model. For this purpose, the importance of the permutation set out in [13] has been used. It can be seen that 8 of the 9 most relevant features only depend on the author’s information and not on the content or information of the tweet. These 9 features are, in order of importance: listed_count, following_count_std, followers_count, tweet_count_std, followers_count_std, quote_count_std, verified, verified_std, tweet_count. In addition, within these characteristics, the information provided by those obtained from the standard deviation of the set of tweets collected for each news item stands out.

The percentage of importance of the most relevant features used in the logistic regression of the hybrid model has also been calculated. To calculate the importance of each feature, f_i , the coefficients of the regression, w_i , have been extracted and the following operation has been carried out $f_i = e^{w_i}$. Finally, the percentage of each of them has been calculated. With this, the most relevant characteristic

for the model, with 10 times more importance over the rest, was the variable that corresponds to the probability returned by the Random Forest that a news item is true using the social information of the news item.

7. Conclusions and Future Work

Throughout the development of this work, it has been observed how the introduction of social information, combined with textual information, has enabled the classification of news, helping to improve the performance of the models. This suggests that, when solving a problem, it would be useful to add social information to the dataset. However, obtaining this information is quite costly both economically and in terms of time.

Additionally, the importance of social features in classifier models has been studied, concluding that author-related features are more important than tweet-related features. The development of a model that combines all textual and social features achieves similar or better results than models that use only textual information.

However, it is crucial to acknowledge several important limitations:

- **Impractical Approach:** Many of the social signals being harvested are post-facto. While disinformation might actually be spreading, many features (such as the number of reposts) would not have stabilized. Thus, while the current approach of augmenting these signals might work post-facto, it is unlikely to work with live data. Even post-facto, it is unclear whether the approach will scale.
- **Flawed Methodology:** The use of balanced training data, and a small set of data at that, is not meaningful. Particularly, it is unclear how learning from such a small corpus would generalize when new kinds of disinformation arise. In practice, the distribution of disinformation-carrying articles compared to genuine ones is far from balanced. Therefore, any realistic methodology needs to incorporate the ability to handle imbalance and transferability from the learning phase. Moreover, adversary behavior might change to emulate the features of good articles or at least stray away from its current behavior, rendering the specific features used for classification obsolete.
- **Too Static and Small Dataset:** The dataset used is too static and small, and lacks adequate diversity to consider any results conclusive. A variety of distinct datasets ought to be used to determine if the ideas actually work in a more general setting.

As a line of future work, it would be a good approach not only to study the individual social metadata of each user, but also to study a social graph of the followers or followers to see the social relationships that exist between them. Additionally, the dataset should be expanded and diversified, and methods should be developed to handle imbalanced data and adapt to changing adversary behavior.

We acknowledge that this work, while preliminary, can trigger useful discussions and provides a foundation upon which more robust and scalable approaches can be built in the future.

Acknowledgments

This work was supported by the HAMiSoN project grant CHIST-ERA-21-OSNEM-002, AEI PCI2022-135026-2 (MCIN/AEI/10.13039/501100011033 and EU “NextGenerationEU”/PRTR).

References

- [1] P. Bharadwaj, Z. Shao, Fake news detection with semantic features and text mining, *International Journal on Natural Language Computing (IJNLC)* Vol 8 (2019).
- [2] R. K. Kaliyar, A. Goswami, P. Narang, Fakebert: Fake news detection in social media with a bert-based deep learning approach, *Multimedia tools and applications* 80 (2021) 11765–11788.

- [3] N. K. Conroy, V. L. Rubin, Y. Chen, Automatic deception detection: Methods for finding fake news, *Proceedings of the association for information science and technology* 52 (2015) 1–4.
- [4] C. Buntain, J. Golbeck, Automatically identifying fake news in popular twitter threads, in: *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, IEEE, 2017, pp. 208–215.
- [5] M. Albahar, A hybrid model for fake news detection: Leveraging news content and user comments in fake news, *IET Information Security* 15 (2021) 169–177.
- [6] N. Ruchansky, S. Seo, Y. Liu, Csi: A hybrid deep model for fake news detection, in: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 797–806.
- [7] K. Shu, S. Wang, H. Liu, Exploiting tri-relationship for fake news detection, *arXiv preprint arXiv:1712.07709* 8 (2017).
- [8] J.-P. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, J. J. M. Escobar, Detection of fake news in a new corpus for the spanish language, *Journal of Intelligent & Fuzzy Systems* 36 (2019) 4869–4876.
- [9] H. Gómez-Adorno, J. P. Posadas-Durán, G. B. Enguix, C. P. Capetillo, Overview of fakedes at iberlef 2021: Fake news detection in spanish shared task, *Procesamiento del Lenguaje Natural* 67 (2021) 223–231.
- [10] J. Canete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, *Pml4dc at iclr 2020* (2020) 1–10.
- [11] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: *PML4DC at ICLR 2020*, 2020.
- [12] F. Barbieri, L. Espinosa-Anke, J. Camacho-Collados, Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond, *Proceedings of the LREC, Marseille, France* (2022) 20–25.
- [13] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, *Bioinformatics* 26 (2010) 1340–1347.