FakeSpreadersWhatsApp.BR: Misinformation Spreaders Detection in Brazilian Portuguese WhatsApp Messages

Lucas Cabral, Diogo Martins, José Maria Monteiro, Javam Machado and Wellington Franco
Universidade Federal do Ceará, Brazil

Keywords: Datasets, Machine Learning, Misinformation Spreaders Detection.

Abstract: In the past few years, the large-scale dissemination of misinformation through social media has become a critical issue. In many developing countries such as Brazil, India, and Mexico, one of the primary sources of misinformation is the messaging application WhatsApp. Recently, a few methods for automatic misinformation detection for the WhatsApp platform were proposed. On the other hand, identifying users who spread fake news is another key aspect of mitigating misinformation dissemination effectively. However, features to describe users on the WhatsApp platform were not found in the literature. This paper proposes a set of 23 features and two approaches (a supervised and another unsupervised) to identify possible misinformation spreaders on WhatsApp. Our results indicate that the proposed features can be used to distinguish between potential misinformation spreaders and users who share credible information with a *F1 Score* of 0.923.

1 INTRODUCTION

In recent years, the large-scale dissemination of misinformation through social media has become a critical problem, undermining public health, social stability and even democracy. In many developing countries such as Brazil, India and Mexico, the WhatsApp messaging app is one of the main sources of misinformation (Martins et al., 2021; Martins et al., 2022). In this context, identifying users who spread false content on this platform is a central task in the fight against disinformation since it allows identifying its origin or its main spreaders. The misinformation spreaders detection task aims to identify malicious users responsible for spreading misinformation on a large scale (Morais and Digiampietri, 2022). The identification of misinformation spreaders makes it possible to create mechanisms whose purpose is to block the misinformation flow, mitigating its dissemination. However, proposals of features to describe users on the WhatsApp platform were not found in the literature.

This work brings three important contributions:

(i) A set of 23 features to describe the behavior of users on WhatsApp, which can be used in the automatic detection of misinformation spreaders.

(ii) Two distinct approaches to identify possible spreaders of misinformation on WhatsApp: thresholding and logistic regression.

(iii) A large-scale, labeled, and public dataset of misinformation spreaders on WhatsApp platform. This dataset, called *FakeSpreadersWhatsApp.Br*, contains 5,364 instances, where each instance represents a user, and 23 different features, collected from public chat groups, using the platform proposed by (de Sá et al., 2021).

Our results indicate that the proposed approaches and features can be effectively used to distinguish between potential misinformation spreaders and users who share reliable information on WhatsApp. The thresholding approach (an unsupervised method) obtained a *F1 Score* of 0.840. The approach based on logistic regression (a supervised method) presented a *F1 Score* of 0.923. Then, we hope that this paper can help researchers understand Brazil's misinformation propagation. The presented ideas can also be used to build misinformation detection systems, which aim to assist users in detecting and filtering out deceptive news.

The remainder of this paper is organized as follows. Section 2 presents our "misinformation spreader" definition. Section 3 discuss the main related work. Section 4 describes the methodology used in this investigation. Section 4 presents the experimental results. Conclusions and future work are presented in Section 5.

Cabral, L., Martins, D., Monteiro, J., Machado, J. and Franco, W.
FakeSpreadersWhatsApp.BR: Misinformation Spreaders Detection in Brazilian Portuguese WhatsApp Messages.
DOI: 10.5220/0011985500003467
In *Proceedings of the 25th International Conference on Enterprise Information Systems (ICEIS 2023) - Volume 1*, pages 219-228
ISBN: 978-989-758-648-4; ISSN: 2184-4992
Copyright © 2023 by SCITEPRESS – Science and Technology Publications, Lda. Under CC license (CC BY-NC-ND 4.0)

2 FUNDAMENTAL CONCEPTS

In this work, we propose an embracing definition for the concept of “misinformation spreader”, which is formulated next:

Theorem 2.1. *Let be a user $u = \{U_u, E_u\}$ of a social network N, who has associated with him a set $U_u = \{u_1, u_2, ..., u_m\}$ of other m users with which u has a connection, a set of engagements $E_u = \{e_1^u, e_2^u, ..., e_n^u\}$, where each $e_i^u = \{p_i, a, t\}$ represents an engagement of u with the publication p_i, for the action a, in the time t. Let the function $Q(s) \mapsto [0,1]$ be a misinformation score assigned to u and τ be a decision threshold. Detecting misinformation spreaders is the task of learning a prediction function $G(u,\tau) \mapsto [0,1]$, satisfying:*

$$G(u,\tau) = \begin{cases} 1 \textit{ (is a misinf. spreader), if} & Q(s) \geq \tau \\ 0 \textit{ (is not a misinf. spreader), if} & Q(s) < \tau \end{cases}$$

A specific definition to categorize a user as a misinformation spreader may vary according to the analyzed social network or the particular behavior one wants to detect. However, it should be considered that the user posts or shares misinformation with unusual frequency or proportion compared to other users of this social network. That is, a misinformation spreader publishes a high amount of misleading publications, or most of his publications contain false information. It is not, therefore, a gullible user who has regular activities on the social network and eventually publishes unreliable information, but users engaged in abnormally disseminating misinformation compared to regular users. It is essential to highlight that, depending on the social network, this behavior often violates its community policies.

3 RELATED WORK

The misinformation spreaders detection is still a problem little addressed in the context of the Portuguese language. Most of the existing works address the *bots* detection problem. In (Leite et al., 2020), a set of rules was proposed to describe and classify *bots* on Twitter. The rules are based on the users behavior, and use as input data the number of tweets bookmarked, the index of answered tweets, and the average of retweets. Using a decision tree, users can be classified by these rules. The best result achieved an **AUC** of 0.97 using the dataset collected by (Cresci et al., 2017).

In (Benevenuto et al., 2008), the authors investigate the problem of detecting malicious users (spammers) on the YouTube platform. Users are represented by three groups of features: **user features**, **video features** and **social network features**. User features include the number of videos added to YouTube, number of friends, number of videos watched, number of videos added as favorites, number of response videos sent and received, number of subscriptions, number of subscribers, and the maximum number of videos added in a day. Video features include the videos length average, number of views, ratings, comments, favorites, honorable mentions, and external links on posted videos. Social network features include clustering coefficient, user rank, betweenness, reciprocity and assortativity. Using these features, an *F1 Score* of 0.81 was obtained in the malicious user detection task.

The effectiveness of the most popular classifiers, such as Random Forest and AdaBoost, in detecting *bots* was evaluated in (Morais and Digiampietri, 2022). The obtained results pointed to the degradation of the efficiency of the classifiers when exposed to new datasets, different from the dataset used during the model training. This result derives, among other factors, from the dependence on information based on the user’s profile, which are frequently changed by *bots* developers whenever they realize that certain features are being used by the detection algorithms.

In (Shahid et al., 2022), the authors provided a comprehensive survey of the state of art methods for detecting malicious users and bots based on different features. In (Rath et al., 2021), the authors presented SCARLET (truSt andCredibility bAsed gRaph neuraLnEtwork model using aTtention), a model to predict misinformation spreaders on Twitter. Using real world Twitter datasets, they show that SCARLET is able to predict false information spreaders with an accuracy of over 87%. In (Rath and Srivastava, 2022), the authors proposed a framework based on a complementary approach to false information mitigation on Twitter inspired from the domain of Epidemiology, where false information is analogous to infection, social network is analogous to population and likelihood of people believing an information is analogous to their vulnerability to infection.

In (Heidari et al., 2021), the authors analyzed sentiment features and their effect on the accuracy of machine learning models for social media bot detection on Twitter. A new set of sentiment features were extracted from tweet’s text and used to train bot detection models. Besides, they proposed a new model for the Dutch language and achieve more than 87% accuracy for the Dutch tweets based on new sentiment features.

4 METHODOLOGY

Regarding the research methodology, the nature of this research is applied. As for the approach, this research is quantitative, organized through the following methodological procedures: (i) extraction of data referring to messages sent by users in WhatsApp public groups, (ii) controlled experiments in the laboratory, and (iii) modeling/simulation for the construction of classifiers based on supervised and unsupervised learning. It is important to highlight that this project seeks to identify misinformation spreaders on the WhatsApp platform, considering features that are independent of the used language.

4.1 The FakeSpreadersWhatsApp.Br Dataset

In this work, we start from the *FakeWhatsApp.Br* (Cabral et al., 2021) dataset, where each line represents a message that a particular user sent in a specific WhatsApp public group. The *FakeWhatsApp.Br* dataset has 282,601 messages sent by 5,364 users, in 59 public groups, between July and November 2018, corresponding to the Brazilian election campaign period. The columns of the *FakeWhatsApp.Br* dataset are the date and time that the message was sent, the sender's phone number, the international phone code, the Brazilian state (if the user is from Brazil), the content (text) of the message, the number of words, the amount of characters, and whether the message contained media such as audio, image, or video. However, the *FakeWhatsApp.Br* dataset does not have any media files. Furthermore, the authors computed how often the same message (with the exact same text) appears in the dataset. A message was considered viral if it was observed more than once in the dataset. For this, only viral messages with identical textual content and more than five words were considered to filter out common messages such as greetings. This subset of the *FakeWhatsApp.Br* dataset contains 6,926 viral messages. Additionally, the messages were anonymized in order to remove personal information such as identity document number, individual taxpayer identification number, zip code, and telephone number, among others. Finally, the messages were manually labeled (as misinformation or not misinformation). The strategy used to build the *FakeWhatsApp.Br* dataset was described in (Cabral et al., 2021). The corpus of *FakeWhatsApp.Br* is publicly available in a repository *online*[1].

From *FakeWhatsApp.Br* we built a new dataset called *FakeSpreadersWhatsApp.Br*, containing 5,364 instances, where each instance represents a user, with features calculated from the original dataset and describing their behavior in the groups during the observed period. For each user, 23 different features were computed, organized into two large groups: activity and network features. Next, we detail each of these features.

4.1.1 Activity Features

As the name implies, activity features quantify the actions taken by users in observed groups. Activity features can be organized into three subgroups: count, proportion, and temporal activity.

The count features are groups (number of groups the user joins), total messages (total number of messages sent by the user), texts (number of text messages sent), media (number of media messages sent), viral (number of viral messages sent), repeated messages (number of repeated messages sent by the user), and misinformation (number of messages labeled as misinformation sent by the user). High values in count features indicate that the user was very active in the observed scope. In addition, high values in the attributes media, viral, and repeated messages indicate propagandist behavior. High values in the misinformation feature indicate misinformation behavior.

Table 1 shows statistical measures that describe the count features in the *FakeSpreadersWhatsApp.Br* dataset. We can see that spreading misinformation is not common among users, as 75% of users sent less than one message containing misinformation. It is also noted that most users do not actively participate in the groups. All count features have distributions with a high concentration of lower values, but with a high standard deviation and a high maximum value, characterizing themselves as long-tail distributions.

The proportion features are: texts (ratio between the number of text messages and the total number of messages sent by the user), media (ratio between the number of messages containing media and the total number of messages sent by the user), viral (ratio between the number of viral messages and the total number of messages sent by the user), repeated messages (ratio between the number of repeated messages and the total number of messages sent by the user) and misinformation (ratio between the number of messages labeled as misinformation and the total number of messages sent by the user). These features seek to capture the relationship between a specific type of message and the total number of messages sent by the user.

Table 2 shows statistical measures that describe the proportion features in the *FakeSpreadersWhat-*

[1] https://github.com/cabrau/FakeWhatsApp.Br

Table 1: Statistical measures of the count-type activity features.

	Groups	Total messages	Texts	Media	Viral	Repeated messages	Misinformation
average	1.16	52.68	29.13	23.55	3.89	2.57	2.13
standard deviation	0.65	138.06	89.74	63.19	15.01	16.26	7.33
min	1.00	1.00	0.00	0.00	0.00	0.00	0.00
Q1	1.00	3.00	2.00	1.00	0.00	0.00	0.00
median	1.00	13.00	6.00	4.00	0.00	0.00	0.00
Q3	1.00	45.00	23.00	19.00	2.00	1.00	1.00
max	11.00	4396.00	3742.00	1360.00	564.00	609.00	147.00

sApp.Br dataset. We have observed empirically that some of these features indicate unusual activity, especially when users have not been very active, but had high values for media and viral features. To illustrate this point with a real example, a user who shared 17 viral messages might not be relevant when looking at the raw amount alone. However, it was observed that 100% of these messages are viral, which is a behavior that does not correspond to that of a regular user. This user is not using the application for conversations, but as a propagandist, just passing on content. The misinformation feature could only be obtained because the data went through a previous manual labeling process. However, the other features can be obtained from an unlabeled dataset.

Temporal activity features are: active days (number of days the user has sent messages), mean, standard deviation, median, and maximum number of messages per day. Table 3 shows statistical measures that describe temporal activity features in the *FakeSpreadersWhatsApp.Br* dataset. Temporal activity features describe the user behavior over time and can be extracted without a previous manual labeling process. Suspicious activities include the user with very sharp bursts of activity alternating with days of no activity. A low mean, with a high standard deviation and a high maximum value, can be a solid indicator to identify misinformation spreaders.

4.1.2 Network Features

Modeling user relationships through a network or graph, can provide relevant information about the misinformation flow. In some social networks such as *Twitter* or *Facebook*, there are well-defined connections between users through the relationship of following (*Twitter*) or friendship (*Facebook*). However, on WhatsApp, these connections are not explicit. Thus, we propose modeling the relationships between WhatsApp users in the form of directed and valued graphs, considering the sending of messages in groups. In this modeling, each node represents a user, and we can consider a graph for each type of message: i) messages in general (General Graph), ii) viral messages (Viral Graph), and iii) messages with misinformation (Misinformation Graph). Considering the general graph, where each node represents a user, a directed edge exists between user i and user j if user i sent a message to a group that user j belongs. The weight of this edge is the number of messages sent by user i to that group. If user i and user j jointly participate in k groups, the weight of the edge from i to j will be the sum of the amounts of messages sent by i to these k groups.

Analogous reasoning can be applied to create the viral graph: there is a directed edge between user i and user j if user i sent a viral message to a group which user j belongs, and the weight of this edge is the number of viral messages sent by the user i to that group. The same goes for the misinformation graph. It can be seen that in the three graphs, the number of nodes is the same, varying the number of edges. The count of the number of edges of each type is presented in Table 4. It can be seen that they are large graphs with many connections.

Figure 1 exemplifies the format of the general graph, using a sample of 2,000 users. Besides, for simplification purposes, the weights and direction of the edges are ignored in this representation. Note that there are isolated groups and a *cluster* of users strongly connected in the center. This is because there are engaged users who actively participate in many groups. Figure 2 illustrates this *cluster* in more detail. Observe the existence of users with high centrality, that is, who interacted with several other users and users who interact only with their local group.

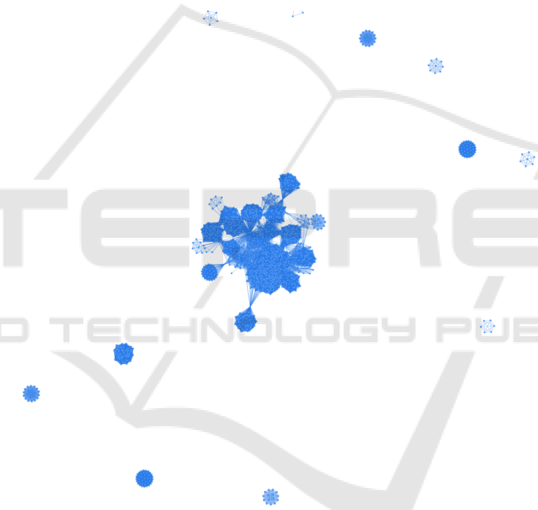
From these graphs, we can obtain some network metrics, which we call here "network features": general centrality degree, general strength, viral centrality degree, viral strength, misinformation centrality degree, and misinformation strength. The last two features can only be computed due to manually assigned labels, while the rest can be computed from unlabeled data. Table 5 shows statistical measures that describe the network features in the *FakeSpreadersWhatsApp.Br* dataset.

Table 2: Statistical measures of proportion-type activity features.

	Proportion				
	Texts	**Media**	**Viral**	**Repeated messages**	**Misinformation**
average	0.567	0.433	0.069	0.039	0.041
standard deviation	0.317	0.317	0.140	0.107	0.107
min	0.000	0.000	0.000	0.000	0.000
Q1	0.333	0.158	0.000	0.000	0.000
median	0.571	0.429	0.000	0.000	0.000
Q3	0.842	0.667	0.085	0.014	0.042
max	1.000	1.000	1.000	0.941	1.000

Table 3: Statistical measures of temporal activity features.

	Active days	**Average daily messages**	**Standard deviation of daily messages**	**Median of daily messages**	**Maximum daily messages**
average	33.3	2.1	2.8	1.4	10.3
standard deviation	30.0	5.0	4.7	4.8	16.9
min	1.0	0.0	0.0	0.0	1.0
Q1	3.0	0.4	0.7	0.0	2.0
median	28.0	1.0	1.4	0.0	4.0
Q3	59.0	2.0	3.2	1.0	12.0
max	120.0	149.0	148.5	149.0	294.0



SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Figure 1: Sample of general graph, with a selection of 2,000 users.

Table 4: Quantities of nodes and edges of the graphs generated to model the relationships between users.

Graphs statistics	
Number of nodes	5.364
Number of edges (General graph)	1.125.326
Number of edges (Viral graph)	551.069
Number of edges (Misinformation graph)	433.204

4.2 Misinformation Spreaders Detection

First, we analyzed the general activity of the 5,364 users of the *FakeSpreadersWhatsApp.Br* dataset by distributing the total number of messages each user sent in the public groups. We observe that the distribution of total messages and other user features has a long tail, with the vast majority of users having low activity. In fact, only 25% of the users sent more than 45 messages. As we are interested in users who had relevant activity, to identify possible misinformation spreaders, we created a clipping containing only users who sent a more significant number of messages than the median, corresponding to 13 messages. This subset contains 2,633 users, called active users. Figure 3 shows the distribution of the number of messages sent by users in the *FakeSpreadersWhatsApp.Br* dataset.

Next, we analyzed the features distributions for this subset of users. As already mentioned, we are interested in the anomalous behavior to define key users, so we used the the well-known *outliers* detection method based on the interquartile range, where

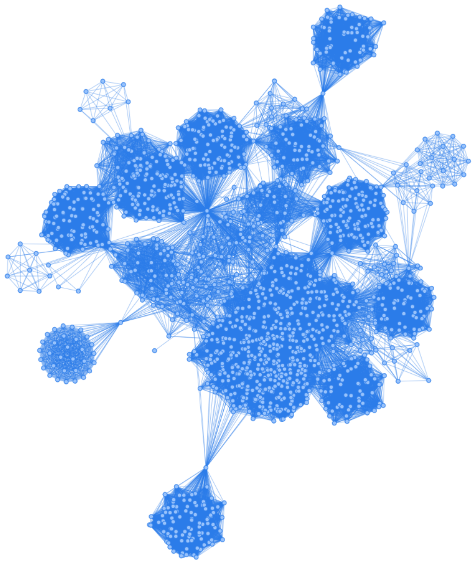


Figure 2: Detail of the graph illustrated in Figure 1, highlighting the strongly connected groups.

Table 5: Statistical measures of network features.

	General centrality degree	General strength	Viral centrality degree	Viral strength	Misinformation centrality degree	Misinformation strength
average	215	10598	105	713	83	386
standard deviation	142	29226	151	2859	136	1392
min	3	7	0	0	0	0
Q1	105	494	0	0	0	0
median	200	2114	0	0	0	0
Q3	278	8322	200	404	153	273
max	1710	672588	1681	96342	1506	28601

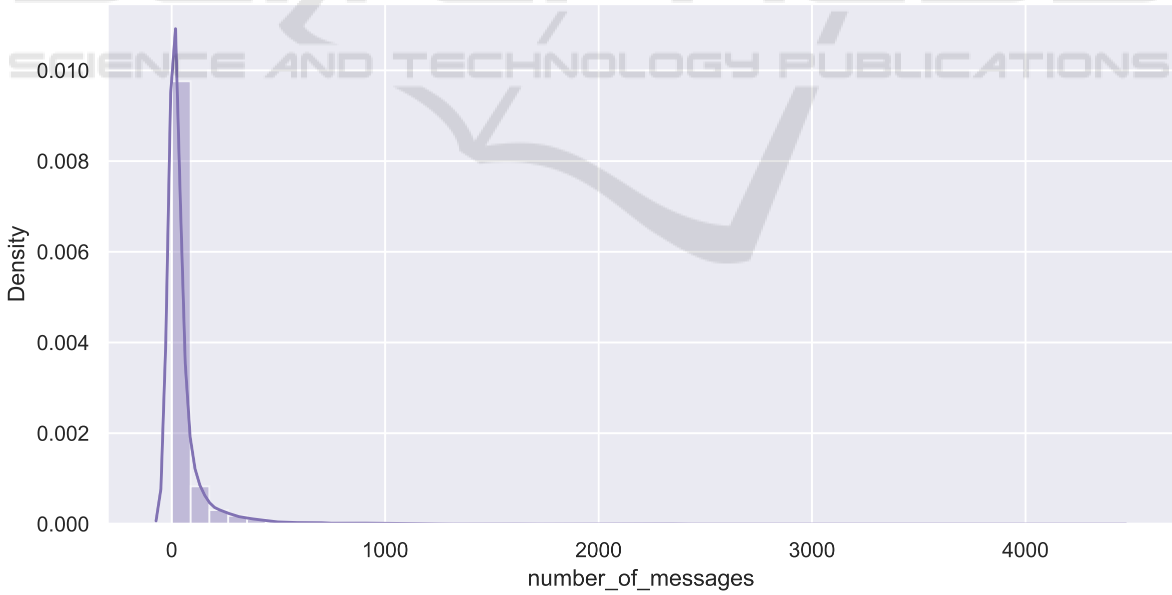


Figure 3: Number of messages sent by users in the *FakeSpreadersWhatsApp.Br* dataset.

an *outlier* is defined as the value equal to $Q3 + 1.5 \cdot IQ$, where $Q3$ is the third quartile and IQ is the interquartile distance of the distribution, considering only the subset of active users.

Then, we propose a definition of misinformation spreader based on the misinformation strength feature. In this work, the misinformation spreaders constitute the group that, among the active users, has an anomalous value for the misinformation strength feature, according to the interquartile distance method. In the *FakeSpreadersWhatsApp.Br* dataset, an *outlier* value at the misinformation strength feature is above the threshold of 28,601. It is important to note that this user category could also have been defined using other feature, such as viral centrality degree, general strength or active days. However, we chose misinformation strength because we are interested not only in the amount of misinformation shared, but in the reach they had, and this feature encapsulates both pieces of information. That is, our definition of misinformation spreader encompasses users who can be characterized as "spreaders" due to the scope and frequency of their actions, which cause more damage than low-range gullible users.

Next, we show four different messages sent by the most active misinformation spreader, following the definition presented previously:

1. • Original Message:
   "*https://youtu.be/iXi3X2XDg6A .... *URGENTE* !! multipliquem este vídeo ao máximo!!*"
   • Message Translated to English:
   "*https://youtu.be/iXi3X2XDg6A .... *URGENT* !! multiply this video to the maximum!!*"
2. • Original Message:
   "*https://youtu.be/WcXXsERafNA. *MAIS UMA FAKE NEWS do HADDAD DESMACARADA!!!* *COMPARTILHEM com todos os seus contatos!!!* vamos colocar este vídeo *EM ALTA* no YouTube!!!!*"
   • Message Translated to English:
   "*https://youtu.be/WcXXsERafNA. *ANOTHER FAKE NEWS from HADDAD UNMASKED!!!* *SHARE this with all your contacts!!!* we'll put this video *UP* on YouTube!!!!*"
3. • Original Message:
   "*Mais uma fake News da mídia........o assassinato do capoeirista não teve nada a ver com política ou muito menos com apoiador de Bolsonaro....... *CANALHAS!! ..... Divulgue este vídeo para todos os seus contatos e grupos do WhatsApp**"
   • Message Translated to English:
   "*Another fake news from the media........the murder of the capoeirista had nothing to do with politics or much less with a Bolsonaro supporter...... *SCAMPS!! ..... Share this video to all your WhatsApp contacts and groups**"
4. • Original Message:
   "**No Ceará, o Comando Vermelho(CV) PROIBIU propaganda de BOLSONARO nos territórios que* *"administra"* *Somente LULA E CIRO Podem. Por serem aliados do CRIME.* Alguém tem dúvida agora da quadrilha?*"
   • Message Translated to English:
   "**In Ceará, the Comando Vermelho (CV) PROHIBITED BOLSONARO's propaganda in the territories it* *"manages"* *Only LULA AND CIRO can. Because they are allies of CRIME.* Does anyone have any doubts now about the gang?*"

Table 6 presents information about misinformation spreaders. One can observe the that the misinformation spreader category is formed by only 2.5% of users, but these are responsible for a large volume of total misinformation, reaching almost 40%. This shows that most misinformation is propagated by a small number of users, whether acting maliciously or not, which reinforces the need to identify these users as a way to mitigate the spread of misinformation.

4.3 Experimental Evaluation

Based on our definition of misinformation spreader presented in the previous section (value of the misinformation strength feature greater than 28,601) and the 23 features of the *FakeSpreadersWhatsApp.Br* dataset, binary classification experiments were carried out to identify whether a user is a misinformation spreader (positive) or not (negative).

It is important to note that the misinformation label can only be assigned due to the message labeling process, which is used to calculate the misinformation strength metric, which defines a user's class. However, there are other features (such as viral centrality degree, general strength and active days) that are known a priori, without any manual labeling process, and these can also be used to identify misinformation spreaders. Nevertheless, exploring these features of misinformation spreaders detection is outside the scope of this paper.

Additionally, it is worth mentioning that the subset formed by active users presents a high imbalance between classes (misinformation spreader and non-spreader). The positive class (misinformation spreader) is the minority, having 132 users, while the negative class (non-spreader) is the majority, having 5,232 users, as illustrated in Figure 4. This imbalance often increases the difficulty of classification, as classifiers may tend to recognize only the majority of class patterns.

Table 6: Description of the misinformation spreader category in terms of number of users, percentage of these users in relation to the total, amount of misinformation sent by users of this category and percentage of misinformation sent by users of this category in relation to total misinformation.

Category	Feature	Threshold	nº of users	% of users	nº of misinformation	% of misinformation
Misinformation Spreader	Misinformation Strength	28,601	132	2.5%	4,533	39.7%

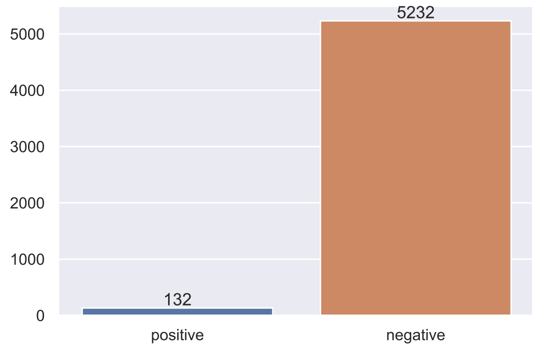


Figure 4: Balancing between user classes. It is perceived that it is a problem of extremely unbalanced classes, where the positive class, of misinformation spreaders, is in the minority.

To evaluate the performance of the two approaches proposed in this work for classifying misinformation spreaders, we performed a random separation of data in the training and test sets in a stratified way, maintaining the proportion between the classes. Thus, the total number of users was split, with 80% for the training set and 20% for the test set. Table 7 presents the amount of data of each class present in each set.

Table 7: Number of negative and positive instances in the training and test sets.

	Training	Test
Positive (misinformation spreaders)	106	26
Negative (non misinformation spreaders)	4,185	1,047
Total	4,291	1,073

4.3.1 Approach 1: Thresholding

This approach assumes a strong correlation between the misinformation strength and viral strength features since all misinformation is also a viral message in the *FakeSpreadersWhatsApp.Br* dataset. In fact, when analyzing the correlation of misinformation strength with other features that can be obtained from unlabeled data, the most strongly correlated variable is viral strength, with a correlation index of 0.87. Thus, a user committed to publicizing and spreading viral messages has a good chance of spreading messages labeled as misinformation. We use this intuition as an approach to detect misinformation spreaders.

The thresholding approach classifies every user with a value greater than or equal to the *outlier* threshold in the viral strength feature as a misinformation spreader. Thus, every user with an anomalous viral strength value, who disseminates viral messages on a large scale, is classified as a misinformation spreader. The threshold observed in the *FakeSpreadersWhatsApp.Br* dataset for the viral strength feature was 5,675. Note that this approach can be used even with unlabeled data. That is, the thresholding approach is an unsupervised method.

4.3.2 Approach 2: Logistic Regression

The second approach proposed in this work consists of using logistic regression. The input is a subset of the 23 user features from the *FakeSpreadersWhatsApp.Br* dataset, normalized by the *z-score* method, using the mean and variance of the training set. This subset is selected using a Decision Tree. To do this, we train the model with the training set and obtain the Gini Importance, which counts the times a feature is used to split a node, weighted by the number of samples it splits. The result of the features' importance is illustrated in Figure 5. Note that the most important feature is, in fact, viral strength. But other features such as the proportion of repeated messages, amount of media, daily messages, and the number of viral messages also add information to the classifier. Thus, we chose the ten most important features: viral strength, amount of media, the proportion of repeated messages, general centrality degree, general strength, average daily messages, amount of viral messages, viral centrality degree, active days, and 95th percentile in the number of daily messages. In addition to the feature selection, another important step is optimizing the decision threshold. Due to the imbalance, the model may tend to estimate lower probabilities for the positive class, so it is necessary to choose an appropriate decision threshold, which in this case, may be less than 0.5. We used the optimal accuracy value in a validation subset, separate from the training set, to optimize the decision threshold choice. In order to avoid inserting noise into the training data, oversampling techniques were not used so that imbalance is dealt with by choosing an appropriate decision threshold.

Note that this approach only can be used if labeled data is available. That is, the logistic regression approach is a supervised method.

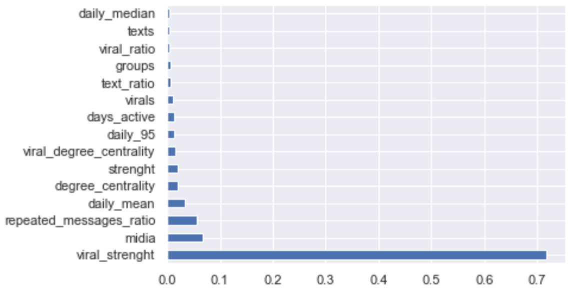


Figure 5: Features Importance.

5 RESULTS

The results obtained in the evaluation are presented in Table 8. Although accuracy is not a suitable metric for this problem, since the classes are very unbalanced, it was also presented. It is observed that the thresholding approach obtained a reasonable result in terms of precision and recall, with a *F1 Score* of 0.840. This approach identified approximately 84% of all misinformation spreaders in the test set.

The results obtained by logistic regression were achieved with a decision threshold of 0.24. That is, the models classifies as misinformation spreader (positive) the user with a probability estimate greater than 24% . The performance obtained by logistic regression was superior in all metrics, in particular in recall, where approximately 92.3% of the misinformation spreaders were identified, and with a high precision, which means a low rate of false positives. Additionally, logistic regression allows interpretability techniques to be applied, which make possible, for example, to understand the contributions of each features to individual predictions. However, logistic regression requires the data to be labeled.

Since viral strength and misinformation strength are strongly correlated, we retrained the logistic regression model without it. In this case, the decision threshold was 0.19. We obtained a *F1 Score* of 0.807, recall of 0.807, and AUC of 0.994. The performance dropped compared to the other methods that rely on the viral strength feature. Even so, the logistic regression approach without the misinformation strength feature identified approximately 87% of all misinformation spreaders in the test set.

Table 8 shows the results of the misinformation spreaders classification using the thresholding and logistic regression approaches. Already Figure 6 illustrates the logistic regression confusion matrix, and

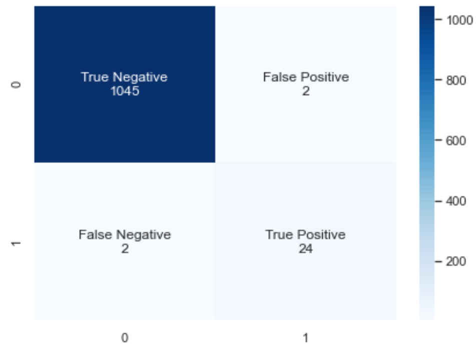


Figure 6: Logistic regression confusion matrix.

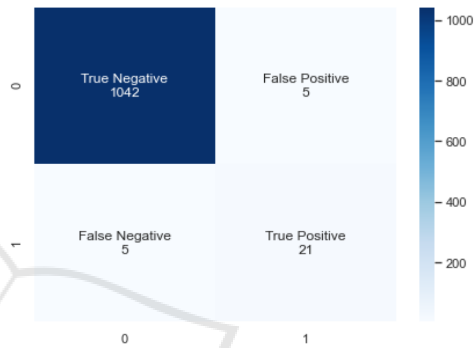


Figure 7: Logistic regression without viral strength confusion matrix.

Figure 7 shows the confusion matrix for the logistic regression trained without the viral strength feature.

Table 8: Results of the misinformation spreaders classification using the thresholding and logistic regression approaches.

Method	ACC	PRE	REC	F1 Score	AUC
Thresholding	0.992	0.875	0.807	0.840	-
Logistic Regression	**0.996**	**0.923**	**0.923**	**0.923**	**0.998**
Logistic Regression w/o Viral Strength	0.990	0.807	0.807	0.807	0.994

6 CONCLUSION

In this paper, we propose a set of 23 features organized into two groups (activity and network attributes) and two distinct approaches to identify possible misinformation spreaders on WhatsApp: thresholding (an unsupervised method) and logistic regression (a supervised method). Our results indicate that the proposed approaches and features can be effectively used to distinguish between potential misinformation spreaders and users who share reliable information on WhatsApp. The thresholding approach obtained a *F1 Score* of 0.840. The approach based on logistic regression presented a *F1 Score* of 0.923. When removing viral strength from the features set, the logistic regression model presented a *F1 Score* of 0.807.

REFERENCES

Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., and Gonçalves, M. (2008). Detectando usuários maliciosos em interações via vídeos no youtube. In *Proceedings of the 14th Brazilian Symposium on Multimedia and the Web*, pages 138–145.

Cabral, L., Monteiro, J. M., da Silva, J. W. F., Mattos, C. L., and Mourao, P. J. C. (2021). Fakewhastapp. br: Nlp and machine learning techniques for misinformation detection in brazilian portuguese whatsapp messages.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2017). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*, pages 963–972.

de Sá, I. C., Monteiro, J. M., da Silva, J. W. F., Medeiros, L. M., Mourao, P. J. C., and da Cunha, L. C. C. (2021). Digital lighthouse: A platform for monitoring public groups in whatsapp.

Heidari, M., Jones, J. H. J., and Uzuner, O. (2021). An empirical study of machine learning algorithms for social media bot detection. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pages 1–5.

Leite, M. A. G. L., Guelpeli, M. V. C., and Santos, C. Q. (2020). Um modelo baseado em regras para a detecção de bots no twitter. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 37–48. SBC.

Martins, A. D. F., Cabral, L., Mourão, P. J. C., Monteiro, J. M., and Machado, J. (2021). Detection of misinformation about covid-19 in brazilian portuguese whatsapp messages. In *International Conference on Applications of Natural Language to Information Systems*, pages 199–206. Springer.

Martins, A. D. F., Monteiro, J. M., and Machado, J. C. (2022). Understanding misinformation about COVID-19 in whatsapp messages. In Chiusano, S., Cerquitelli, T., Wrembel, R., Nørvåg, K., Catania, B., Vargas-Solar, G., and Zumpano, E., editors, *New Trends in Database and Information Systems - ADBIS 2022 Short Papers, Doctoral Consortium and Workshops: DOING, K-GALS, MADEISD, MegaData, SWODCH, Turin, Italy, September 5-8, 2022, Proceedings*, volume 1652 of *Communications in Computer and Information Science*, pages 14–23. Springer.

Morais, D. and Digiampietri, L. A. (2022). Evaluating social bots detection approaches in different domains. In *XVIII Brazilian Symposium on Information Systems*, SBSI, New York, NY, USA. Association for Computing Machinery.

Rath, B., Morales, X., and Srivastava, J. (2021). Scarlet: Explainable attention based graph neural network for fake news spreader prediction. In Karlapalem, K., Cheng, H., Ramakrishnan, N., Agrawal, R. K., Reddy, P. K., Srivastava, J., and Chakraborty, T., editors, *Advances in Knowledge Discovery and Data Mining*, pages 714–727, Cham. Springer International Publishing.

Rath, B. and Srivastava, J. (2022). *Spreader-Centric Fake News Mitigation Framework Based on Epidemiology*, pages 31–54. Springer International Publishing, Cham.

Shahid, W., Li, Y., Staples, D., Amin, G., Hakak, S., and Ghorbani, A. (2022). Are you a cyborg, bot or human?—a survey on detecting fake news spreaders. *IEEE Access*, 10:27069–27083.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS