

Automatic Detection of Textual Disinformation in WhatsApp Messages

A technical and concise approach to fake news classification using the fakeWhatsApp.BR_2018.csv corpus.



Corpus Building Strategy

The base dataset is fakeWhatsAoo.BR_2018.csv, which focuses on messages from public WhatsApp groups. Strict selection and filtering were crucial to the quality of the training.

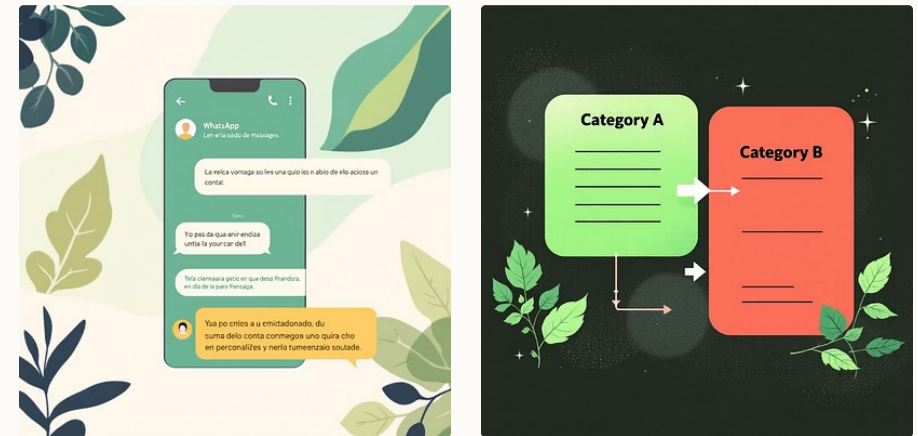
Data Selection

- Filter for viral texts (viral=1).
- Exclusion of media content (media=0).
- Elimination of duplicate and unlabeled texts (-1).

Classes Definition

- Class 1 (Fake): Misinformation texts.
- Class 0 (Real): Verified authentic texts.
- Total of 4588 texts after cleaning.

The final texts for analysis (**X**) are the non-media content from the text column, and the labels (**y**) are the binary values from the misinformation column.



Distinctive Characteristics of Texts

Disinformation and real texts exhibit distinct lexical and structural patterns, which serve as clues for automatic detection.



Misinformation (Fake)

- Excessive use of **emojis** and visual repetition.
- **Imperative** language (call to action, "See explicitly").
- Chain structure, **seeking virality** and **urgency**.



Real Texts (Real)

- Similarity to **everyday conversations**.
- More **natural** and sporadic use of emojis.
- **Normalized** use of punctuation and capitalization (easy to identify intuitively).



Corpus Statistics After Filtering

Table 1 summarizes the final corpus size and the averages of textual metrics, essential for understanding the information density in each class.

Metric	Total Count	Characteres Mean	Words Mean	Shares Mean
Total Corpus	4588	574.50	87.25	11.89

Comparative Statistics by Class

Class	N.º Texts	Words Mean	Shares Mean	Characteres Mean
Misinformation (1)	2041	113.87	4.974	719.40
Real (0)	2547	60.044	3.454	408.92

It is observed that misinformation tends to be longer (more words) and to be shared more, reflecting its design for viralization.



Pre-Processing and Feature Engineering

Preprocessing is essential to transform raw text into data that ML models can interpret effectively, focusing on the most informative words.

Text Cleaning

Inserting spaces for **emojis** and **punctuation**, followed by their removal.

Stopwords Removal

Elimination of high-frequency words (pre-defined and manual) that do not add meaning.

Normalization

Stemming words and replacing URLs with their domains (e.g., www.google.com).

Truncation and Filtering

Ignore patterns like kkk+ and keep only the **first 100 words** of each sentence.

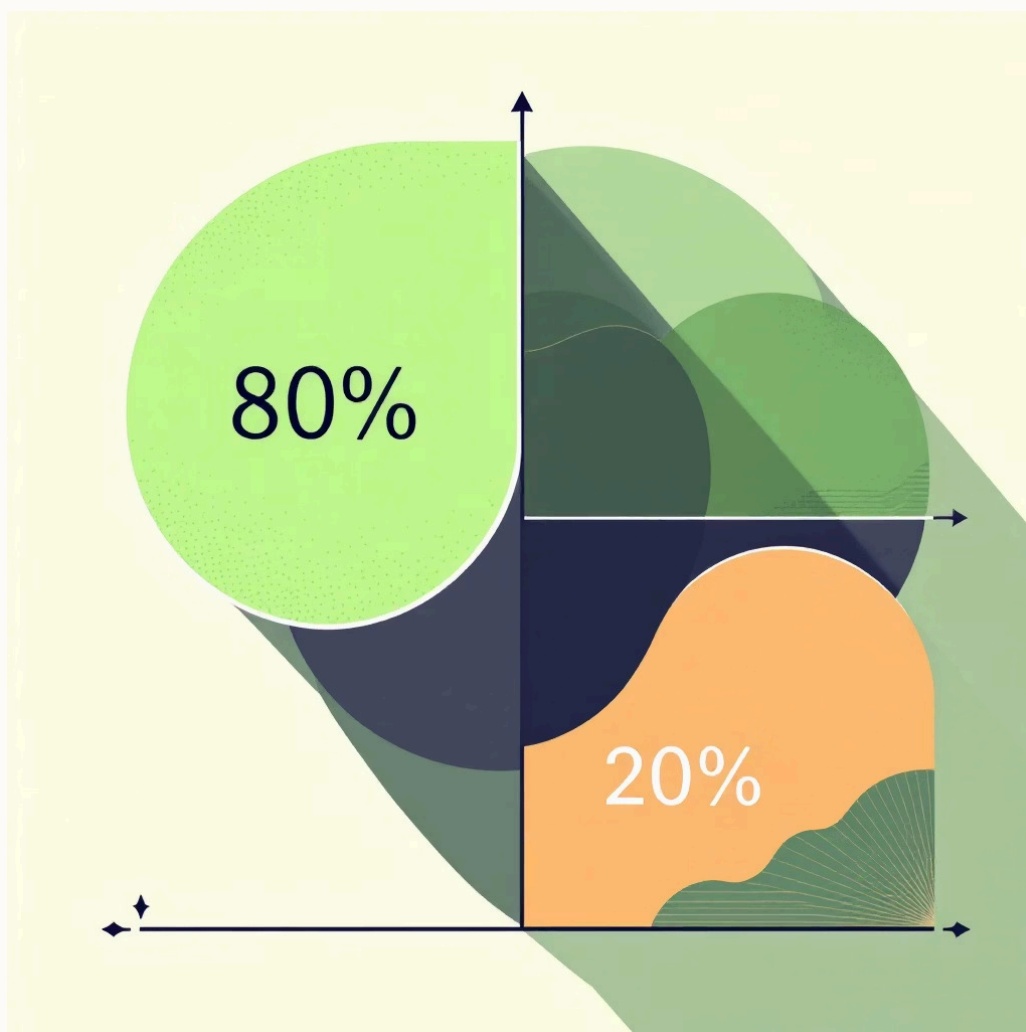
Division and Vectorization Strategy

The dataset division and vectorization technique were defined to ensure robustness and performance of the models.

Dataset Division

- 80% (Training) / 20% (Test) split with `random_state=42` for replicability.

Stratification (stratify=y) was applied to maintain the proportion of classes (Fake/Real) in both sets.



TF-IDF Vectorization

- Using Term Frequency-Inverse Document Frequency (TF-IDF).
- Creating vectors for unigrams, bigrams, and trigrams.
- The vectorizer learns the mapping from the training set only; the test set is transformed using this mapping.



Model Training and Metric Selection

Seven classification models were evaluated, focusing on metrics that prioritize the correct identification of misinformation, avoiding harmful false positives.

7

Tested Models

LR, NB, LSVM, SGD,
SVM, KNN, RF.

1

Max. Recall

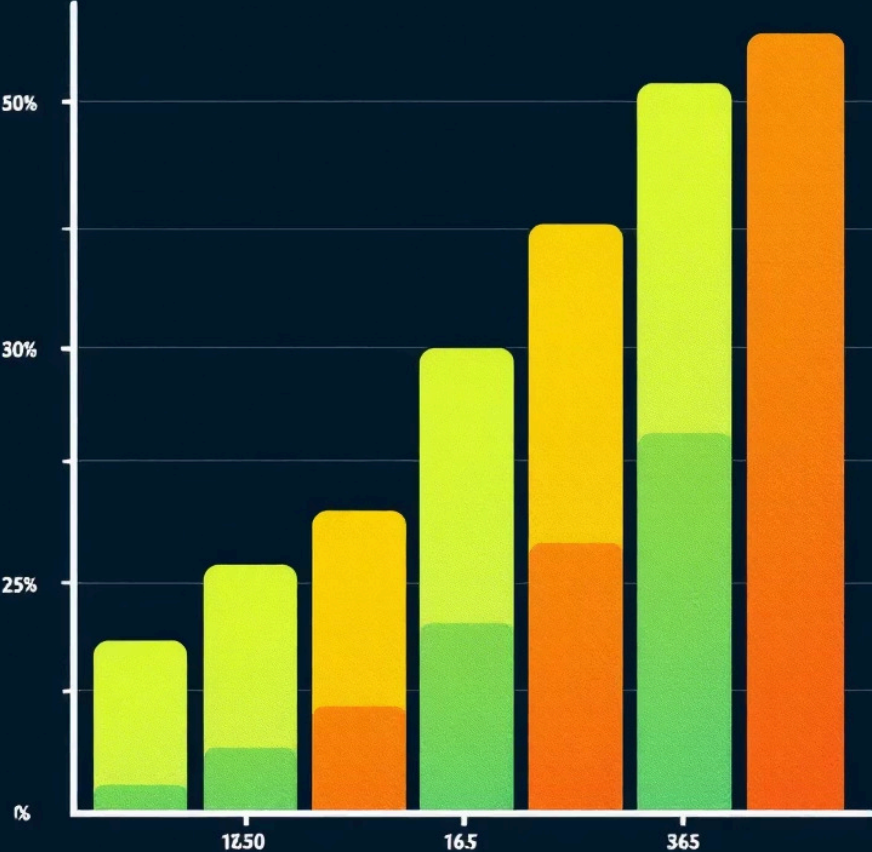
Priority metric:
proportion of
correctly identified
fake messages.

3

Best Models

LSVM, SGD, and NB,
with F1-Score as the
balance metric.

The **F1-Score** was adopted to reconcile **Recall** (detecting fake news) and **Precision** (minimizing the marking of real messages as false, which is highly harmful).



Experiment 1: Weight Assignment Analysis (LIME)

I used LIME (Local Interpretable Model-agnostic Explanations) to compare the keywords that LSVM, SGD, and NB models use to classify a 200-index sentence.

Analyzed Text(Index 200)

"Liguei para o **Diretório** do **PT** oferecendo ajuda.... Acabei **desistindo**. Eita "povinho mal educado!!!"

This sentence presents mixed predictions, making it ideal for model consistency analysis. We selected the four words with the highest combined weight for comparison:

Word/Model	MNB	LSVM	SGD
PT	0.02 (Fake - 4 ^a)	0.10 (Fake - 1 ^a)	0.14 (Fake - 1 ^a)
Diretório	0.04 (Real - 1 ^a)	0.09 (Real - 2 ^a)	0.11 (Real - 2 ^a)
do	0.01 (Real - 1 ^a)	0.03 (Fake - 9 ^a)	0.04 (Fake - 8 ^a)
desistindo	0.01 (Real - 8 ^a)	0.04 (Real - 4 ^a)	0.04 (Real - 9 ^a)

The models agree that **PT** increases the probability of being **Fake** and **Diretório** increases the probability of being **Real**, demonstrating consistency. The word "do" has irrelevant weight, as expected for a pseudo-stopword.

Experiment 2: Global Clue Analysis (LIME)

In this experiment, I use LIME to discover which words contribute most to classifying 300 WhatsApp texts as fake or real, calculating the average and frequency of LIME explanatory words across all models to determine general clues.

The most common words from the top three models were:

- Fake: sentiu/Senhor/suspende/eleitorado/tô
- Real: Maria/padre/Maior/acertou/_____

<div><div></div><div>01100</div></div>				
FAKE				
Word	MNB		LSVM	SGD
sentiu	-0.070213 (5 ^a)		-0.137195 (2 ^a)	-0.17831 (3 ^a)
Senhor	-		-0.102820 (6 ^a)	-0.134738 (8 ^a)
suspende	-0.070602 (4 ^a)		-0.122129 (3 ^a)	-0.183526 (2 ^a)
eleitorado	-0.074198 (3 ^a)		-0.147824 (1 ^a)	-0.197468 (1 ^a)
tô	-0.066830 (6 ^a)		-0.099314 (10 ^a)	-0.125804 (6 ^a)

REAL				
Word	MNB		LSVM	SGD
Maria	0.065470 (3 ^a)		0.102127 (2 ^a)	0.136981 (2 ^a)
padre	0.064031 (5 ^a)		0.076601 (10 ^a)	0.110970 (7 ^a)
Maior	0.052582 (8 ^a)		0.086258 (4 ^a)	0.113512 (5 ^a)
acertou	0.067527 (2 ^a)		0.081617 (5 ^a)	0.112295 (6 ^a)
_____	0.148821 (1 ^a)		0.300264 (1 ^a)	0.332011 (1 ^a)

The overall analysis reveals that the models use similar semantic and contextual cues, with the Real class featuring terms more closely linked to proper nouns or less political/viral themes.

Conclusion and Next Steps

Detecting misinformation requires a balance between the model's ability to capture viral signals (Recall) and accuracy on neutral texts (Precision).



Models Coherence

LSVM, SGD, and NB are the best predictors, showing remarkable similarities in the learned words (LIME cues).



Crucial Clues

Highly politicized or imperative words are strong indicators of misinformation. The average number of words/shares also reinforces this distinction.



Future Improvements

- Apply LIME on examples in which the model fail to classify correctly (fake positives/negative).
- Execute LIME multiple times on the same example and see if the the wieghts are the same.

