Extraction and Visualization of Fake News Indicators

Karan Prem Manghnani
Instituto Superior Técnico
December 2020

ABSTRACT

This dissertation presents FactMe, an online visualization tool that receives a news article and computes a range of metrics intended to provide consumers with information quality indicators. The consumption of news over the internet is rapidly increasing, and becoming increasingly difficult to judge the trustworthiness of news found on the web. Since the web is open, this openness can contribute to spreading disinformation. In the last years, disinformation has caused severe damage and can cause more ahead. This shows how vital can a disinformation detection tool be. Recent advances allow us to analyze news and use natural language processing and machine learning algorithms to detect disinformation and help consumers evaluate the credibility of the news articles. In this work, I have surveyed previous research on content-based (linguistic) and context-based metrics, which helped me develop a learning model using the linguistic indicators that achieved an accuracy of 96% in predicting the articles' veracity. Furthermore, this work provides an assessment from a consumer's perspective, which has shown a positive user satisfaction with an average score of 8.48 in QUIS and at least a small impact of the disinformation indicators on the consumers in predicting the article's veracity.

Author Keywords

Disinformation indicators; Disinformation visualization; Fake news.

INTRODUCTION

The role of traditional information channels, like television and newspapers, on how we collect and consume news is becoming less prominent. The growth of social media and online news sources has played a crucial role in this transformation. In addition, the consumption of fresh news content over the internet is rapidly increasing, resulting from an abundance of technology, which can expose users globally. However, this positive impact comes at a cost; it can also contribute to the spreading of disinformation, which can have many malicious purposes like promoting ideologies, gaining favor in political elections, earning money, and taking revenge, among many other reasons.

Social media has become a critical publishing tool for journalists, and the primary consumption method for citizens looking for the latest news [14]. However, online content proliferation also brings disinformation with it in social media platforms like Facebook, Twitter, and WhatsApp.

Disinformation is false or misleading information spread deliberately to deceive. This is a subset of misinformation, which is also false or inaccurate information. While misinformation is shared regardless of intent to mislead, disinformation is shared deliberately [6]. The terms misinformation and disinformation have often been associated with the term fake news. Fake news is "a news article that is intentionally and verifiably false" [1]. Therefore, it includes news articles intentionally written to mislead or misinform readers, and that can be verified as false through other sources. The authors, who studied the 2016 United States election, noticed a massive amount of fake news websites shared in social media and their impact on the elections. They showed that fake news was very persuasive. The spreading of fake news in social media was very successful and could have misled the population's voting, and so the presidential election. A more recent example is COVID-19 fake news, which has caused much panic among people. According to BBC[1], coronavirus related misinformation may have caused the death of at least 800 people in the first three months. An example of these misleading news states that drinking methanol or alcohol-based cleaning products could cure the virus.

The damage of disinformation leads us to realize how vital a fake news detection tool could be. Recent research has shown techniques and approaches to detect fake news with artificial intelligence tools. An approach for news analysis consists of using NLP for feature extraction [12]. Feature extraction for disinformation analysis can be either content-based or context-based. Content-based features rely on linguistic features, referring to information that can be directly extracted from the text. With statistics on these features, we obtain a structured representation in terms of **linguistic disinformation metrics**. On the other hand, context features are extracted by considering relevant information surrounding the actual social media post or news content. The most used context features refer to the analysis of users, news sources, propagation structures of the information on social media and other users' reactions to the news. These types of features I call as **non-linguistic disinformation metrics**. A variable computed from one or more metrics, called an **indicator**, quantifies disinformation and helps consumers decide the credibility of the news. In general, indicators can be based on content (linguistic analysis) and context (non-linguistic analysis). Multiple online tools

[1] https://www.bbc.com/news/world-53755067

ACM ISBN 978-1-4503-2138-9.
DOI: https://doi.org/10.1145/3313831

like FactMata[2] and NewsGuard[3] have been made using many different approaches to help readers detect and judge online news by presenting indicators.

This dissertation's main purpose was to develop an online visualization tool, named FactMe, for computing a range of metrics (content-based and context-based) intended to provide information quality indicators for news articles in Portuguese. The indicators will empower news consumers to judge the credibility of the articles. FactMe assists users as a Web application that, given a URL of an article or its text, presents to users a range of disinformation indicators in an explainable way, helping them judge the article's credibility.

RELATED WORK

Linguistic disinformation metrics and indicators

Caled and Silva [2] proposed linguistic metrics that can be grouped into six categories of indicators described below:

Affectivity Indicator: Disinformation often employs a higher number of emotional words than informative text [7]. Affectivity metrics compute statistics on emotional words present in text and highlight variations in the informative content. These metrics are based on a three-dimensional emotion representation, proposed by Osgood et al. [10] as the Theory of Emotions:

- **Valence**: Measures how pleasant or unpleasant an emotion may be. For instance, fear is an unpleasant emotion and has a high score on the displeasure scale however, joy is a pleasant emotion.
- **Arousal**: Measures the intensity of the emotion. For instance, anger and rage are both unpleasant emotions, but rage has a higher intensity than anger.
- **Dominance**: Measures the control over specific stimulus. For instance, fear and anger are both unpleasant emotions, but anger is a dominant emotion, while fear is a submissive emotion.

Metrics in this category are calculated using the words found for each dimension considering statistics like minimum, maximum, standard deviation, average, and the difference between minimum and maximum.

Behavioral and Physiological indicator (BP): In a news article, the author also shows the feelings he had about the reported information. Such feelings, which depend on his behavior and psychological processes, can help detect disinformation [13]. The BP category corresponds to six processes such as: **Biological**; **Cognitive**; **Perceptual; Personal**; **Relative**; **Social**.

The BP metrics can be calculated as the fraction of the BP words in each category relative to the total number of words in the text.

[2]https://factmata.com/

[3]https://www.newsguardtech.com/

Emotion indicator: One characteristic of disinformation is that it may make an inflammatory emotional appeal to the reader [7]. It is then essential to detect the emotions that the news transmits. This category is focused on theoretical models for discrete emotion based on the six basic emotions of Ekman [5]: anger, disgust, fear, happiness, sadness, and surprise.

The metrics in this category can be calculated as the fraction of each category's emotion words relative to the total number of words in a text.

Grammatical indicator: This category is based in PoS tagging considering the role, definition, and context of the terms. Grammatical metrics consider content words (i.e., nouns, verbs, adjectives, adverbs) and function words (i.e., prepositions, pronouns, conjunctions, determiners). This category also includes the degree of informality (or typographical error ratio) of a text.

Zhou et al [13] and Carvalho et al. [3] proposed the following metrics in this category:

1. Content diversity: $\frac{\text{number of distinct content words}}{\text{total number of content words}}$
2. Expressivity: $\frac{\text{sum of occurrences of adjectives and adverbs}}{\text{sum of occurrences of nouns and verbs}}$
3. Informality $\frac{\text{number of misspelled words}}{\text{total number of words}}$
4. Modifiers ratio $\frac{\text{number of modifiers (adjectives and adverbs)}}{\text{total number of words}}$
5. Non-immediacy: $\frac{\text{number of 1st and 2nd pronouns}}{\text{total number of words}}$
6. Pausality $\frac{\text{number of punctuation signals}}{\text{total number of sentences}}$
7. Content word representativeness (PoS): $\frac{\text{each content word}}{\text{remaining words in text}}$
8. Redundancy: $\frac{\text{number of function words}}{\text{total number of sentences}}$

Sentiment polarity indicator: The sentiment is expressed through subjective expressions to describe people's opinions, appraisals, or feelings toward a given target [8].

Carvalho et al. [3] proposed the following metrics for this category:

1. Polarity information: $\frac{\text{positive and negative words}}{\text{number of words}}$
2. Polarity contrast: number of sequences where negative words follow positive words and vice versa.

Subjectivity indicator: This category is responsible for distinguishing factual information from subjective information. Subjective expressions are used to express an opinion, emotion, evaluation, stance, or speculation. Subjective terms could be strong subjective (terms that are seldom used without a subjective meaning) and weak subjective (terms that commonly have both subjective and objective uses) [11]. The metrics proposed by the authors are calculated as the fraction of strong or weak subjective words relative to the total number of words in a text.

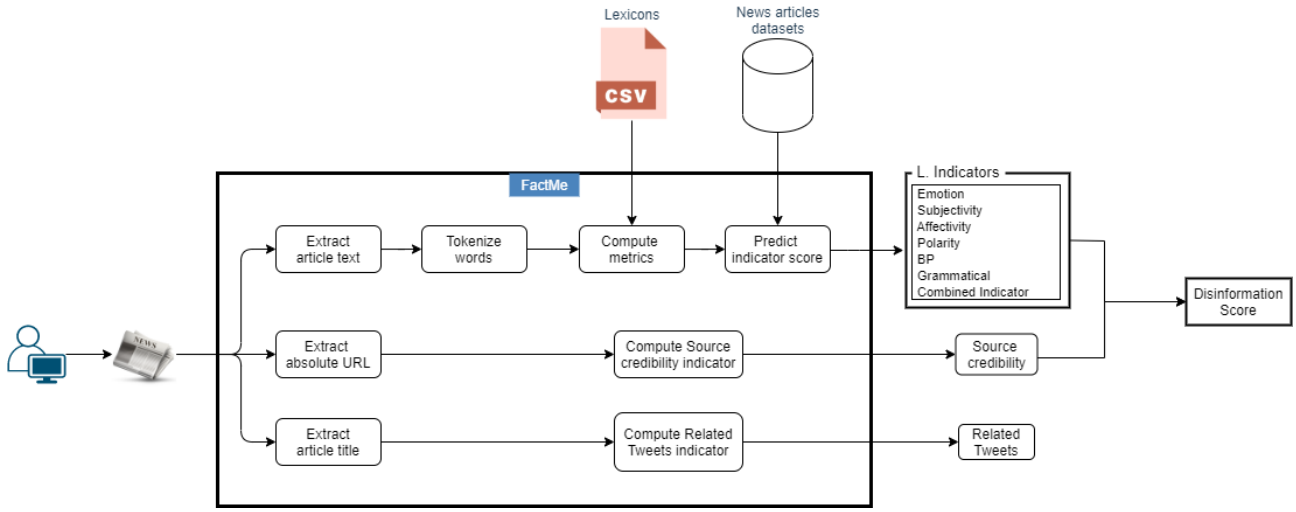

Figure 1. Information processing of FactMe.

FACTME

FactMe is a web application that computes a range of metrics (content-based and context-based) intended to provide information quality indicators for news articles in Portuguese. FactMe receives a news article and presents indicators of disinformation to users in an explainable way to empower news consumers to judge the credibility of the articles. Figure 1 shows how FactMe processes information under the following steps:

- The process starts when the user gives the URL of an article as input. FactMe extracts the article text, the absolute URL, and the article title from the website. Alternatively, if the user gives the article text as input, the extraction phase is skipped.
- The absolute URL is used to identify the source and later compute its credibility. If no URL is provided, the source credibility indicator will show as a not trusted source.
- The title is used to identify related tweets and later compute a disinformation indicator from that information. If no URL is provided, the user can also provide the title along with the article text to identify related tweets. If no title is provided, the related tweets indicator will not search for any tweets.
- The text of the article goes through a tokenization process.
- After tokenization, FactMe computes disinformation metrics for the news article based on their respective lexica.
- Using the computed metrics, an indicator score is predicted for each category of indicators along with a combined indicator.
- Using the combined indicator and the source credibility indicator, a disinformation score is finally computed.

FactMe user interface

The homepage offers two alternatives for the article input. Users can provide a URL of a news website. This method uses Newspaper3k[4], a Python library, to scrap and extract news articles from their website. Alternatively, as this option may not extract the website's article correctly, the user can provide the article's text. This method is also useful to give as input articles extracted from other sources, e.g., social media.

After submitting the article, users obtain disinformation analysis results. Figure 1 shows the interface of a search result in FactMe. The interface includes the URL of the article, the extracted text from the article's website, the linguistic indicators and their respective score, the context indicators, and a disinformation score. The disinformation score will present the article's judgment represented in a color gradient scale with three colors (green, yellow, or red). The indicator's score corresponds to the likelihood of the article being fake for each indicator and is represented in a scale bar divided into three sections, each one having one color, green (score <= 33%), yellow (33% < score <= 66%), and red (score > 66%). The colors of the score have the following meaning:

- **Green**- Shows no perceived or irrelevant misleading tendencies in the news. The user can have a high level of trust in the article.
- **Yellow**- This reveals that the article is classified as dubious because some fake tendencies have been detected.
- **Red**- This shows the user that high fake tendencies were detected, and the user should read the article with caution.

Each indicator has a detail button. For the linguistic indicators, the detail button describes each one, showing all the metrics computed for the specific indicator and the number of words found for the specific metric. The source credibility indicator presents the description, the credibility label, and the official website of the source if trustable. The related tweets indicator shows the tweets found for the specific article, the user, the retweets likes, followers, and verifiability of the user.

[4]https://newspaper.readthedocs.io/en/latest/

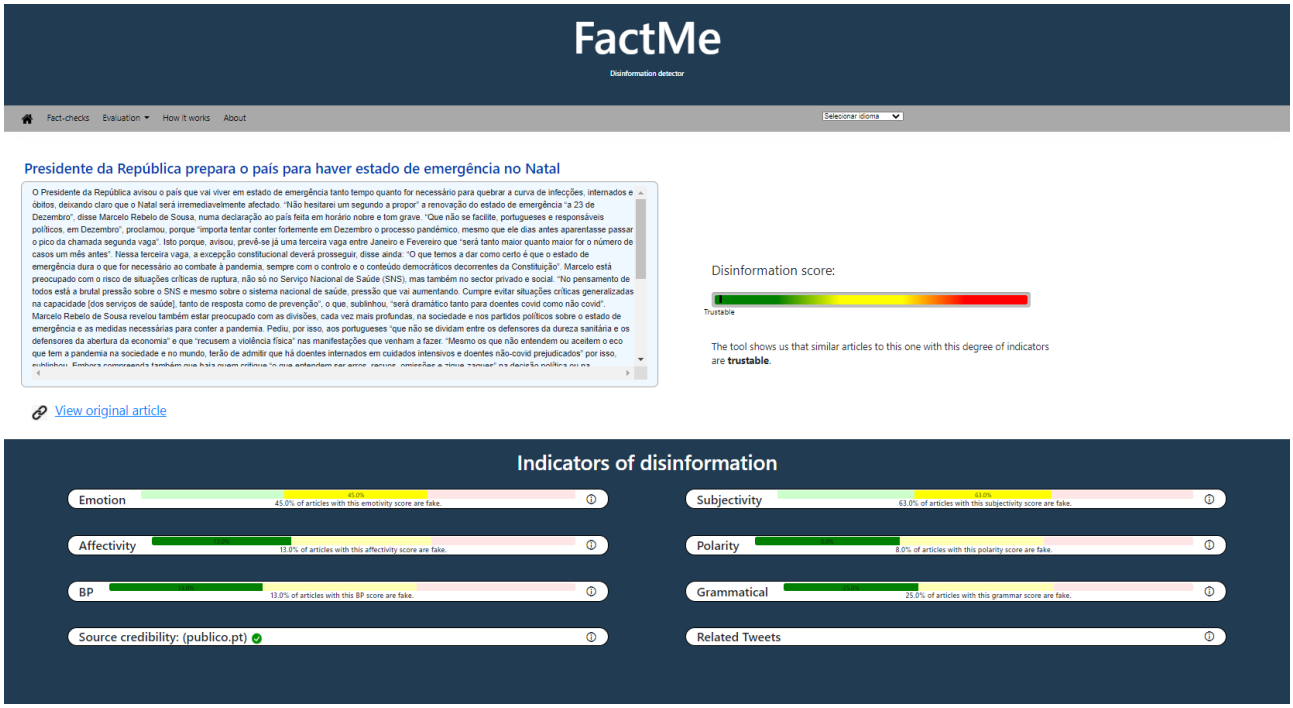

Figure 2. User interface presenting indicators of disinformation for a news article.

The FactMe also includes the following navigational components:

- **Fact-checks**- This component shows the articles that have been submitted to FactMe and their classification. The users can check the last articles submitted by the consumers without the computational time since it is saved in the Firebase[5] (NoSQL database program, which stores data in JSON-like documents).
- **How it works**- This component explains how FactMe works, including a video demonstrating the process to submit, a detailed description of the indicators, and an explanation of the classification.
- **About**- This component explains the goal of FactMe.

Computation of metrics and indicators

Computation of linguistic metrics

The data used to analyze and to compute the linguistic metrics is taken directly from an article text. The pre-processing includes the following steps: (i) text is split into sentences; (ii) tokenization is performed; (iii) punctuation is removed, and (iv) words are lowercased. Prof. Danielle Caled provided a Python library that receives as input a list of words and computes the metrics presented in *Linguistic disinformation metrics and indicators* section using the respective lexica of each category of indicators that were presented in this section. To reduce the time complexity, the lexica were serialized and converted into a byte stream, and deserialized whenever necessary. This library was used and adapted to receive the article tokens that were retrieved from the article text. With this computation, it was possible to calculate all the metrics for each category of linguistic indicators.

Computation of context indicators

- **Source credibility** - This indicator uses the ERC (Entidade Reguladora para a Comunicação Social)[6], which is the entity responsible for the regulation and supervision of all entities that pursue social communication activities in Portugal. The ERC keeps an excel sheet (*Listagem de Publicações Periódicas*), which contains all the news agencies registered in Portugal and their information like workers, website, location, editor, Etc. To be registered by this entity, the journalistic company must be classified as a press by the Portuguese law of press (Law number º 2/99, on 13th January), which requires the company to verify the news's veracity. Being registered by this entity makes a source more credible, and so the article. FactMe obtains the absolute URL using the urllib library in Python, then it analyzes if the URL appears in the sheet, and it shows to the user if the information source is registered or not by ERC.

- **Related Tweets**- A fake article can be shared in a viral way across the internet, and one of these platforms is Twitter. To show to the users how the submitted article is being shared on Twitter, FactMe takes into account some properties like the number of users who shared the article, number of retweets each tweet had, number of followers, number of likes, date and if it is a verified profile by Twitter. To search the tweets, I used Tweepy[7], a Python library that accesses Twitter API. I used the search method from Tweepy

[5]https://firebase.google.com/

[6]https://www.erc.pt/pt/listagem-registos-na-erc

[7]https://www.tweepy.org/

that receives the title of the article as a query, the maximum number of return tweets as 100, only tweets written in Portuguese are retrieved, and I also used the extended tweet mode to allow to get the full text of the tweet. Firstly, I attempted the OR approach of the Twitter API (which searches each word individually) by searching the words of the articles' title without the stopwords. This method was not effective because the results did not correspond to the search made, retrieving tweets that are not related to the article context. Another approach I adopted was to use Tweepy to search for the full title of the article. This approach shows all the Twitter users that shared this article, showing to end-user the tweet properties established by FactMe and allowing them to analyze how the article is being shared on Twitter.

Calibration of linguistic indicators

To evaluate the feasibility of the linguistic metrics of each category of indicator, I collected two datasets that contain news articles, Polígrafo[8], and Fake.br Corpus [9], to compute the disinformation metrics. The datasets are described as follows:

- **Polígrafo** is a Portuguese online journalistic project that does fact-checking to Portuguese news articles. The classification of the information is made with the following five rating scales: (1) True, (2) True, but; (3) Inaccurate; (4) False; (5) Pepper on the tongue. I gathered 50 text articles from various sources that were classified by Polígrafo and labeled as True, articles that were classified as "True" and "True, but" and labeled as False, the articles classified as "False" and "Pepper on the tongue".

- **Fake.Br.Corpus**- This corpus is composed of manually labeled news written in Brazilian Portuguese (3,600 true + 3,600 false). The Fake.Br Corpus contains news articles published from January 2016 to January 2018. I also considered this dataset due to the large amount of news it contains in Portuguese.

I created a dictionary of metrics with the computed metrics from Polígrafo and Fake.br Corpus, where I labeled each article with respect to its veracity (true or false), assigned the corresponding metrics, and evaluated the metrics in the following two ways:

1. Creating a box plot of metrics for each category of indicator using matplotlib[9] (plotting library for Python) to compare the distributions of true and fake labels between the two datasets.

2. Analyzing the disinformation prediction as a classification problem using Logistic Regression and evaluating each indicator's performance using the respective metrics.

Performance of linguistic indicators

To evaluate the performance in disinformation detection with machine learning tasks, I chose Logistic Regression due to its simplicity, speed, and reliability for binary classification. The code used the Logistic Regression function in sklearn library[10] with the parameter max_iter as 8000 and solver as "saga" due to the large dataset. All other parameters are set as default. The datasets were split into stratified training/testing (0.20/0.80). I used as features the metrics of each indicator individually to evaluate each category of indicators, and also used as features all the metrics gathered, which I called the **combined indicator**. I ran the Logistic Regression algorithm for the dictionary of metrics that were computed from both datasets and used to create the boxplots, reporting the precision, recall, F1 score metrics, and accuracy for both classes (true and false).

Table 1(a) presents the indicator's performance for the Polígrafo dataset. We can see that subjectivity, polarity, BP and grammatical metrics had low accuracy, not detecting most of the article's veracity correctly. However, emotion metrics produced a satisfactory accuracy (67%), and even better was affectivity accuracy (83%), predicting the majority of the labels. The combined indicator showed an accuracy of 67%. Table 1(b) presents the performance for the Fake.br Corpus dataset. We can see that subjectivity metrics produced the lowest accuracy but still greater than Polígrafo, and emotion metrics had an accuracy of 60%, which was lower than Polígrafo. Polarity had good accuracy with 76%. The best results were from grammatical (90%) and affectivity (93%), which predicted the majority of the articles' veracity correctly. The combined indicator had very good accuracy (96%), predicting the article's veracity very well.

Discussion

First, I analyzed the boxplots of the metrics computed for the Polígrafo and Fake.br Corpus datasets in order to differentiate true and false articles. We saw that emotion metrics differentiated better the articles' veracity in Polígrafo, and affectivity and polarity metrics differentiated better in Fake.br Corpus. In the second analysis, we saw that only the emotion indicator had better accuracy in the Polígrafo dataset; all other metrics were better in predicting the article's veracity in Fake.br Corpus. Also, the combined indicator had much better accuracy in the Fake.br Corpus dataset. In general, the result of the Fake.br Corpus classification, when compared to Polígrafo's classification results, are better in both, single categories of indicators and the combined indicator. The only metrics that gave us better results in Polígrafo was from the emotion indicator, which will be used for prediction using the Polígrafo dataset. All other metrics and the combined indicator will be used from the Fake.br Corpus in disinformation tasks.

Computation of the linguistic indicator and disinformation scores in FactMe

Disinformation detection is formulated as a classification task to compute the score of the linguistic indicators. The input is defined by the metrics computed from the news article given by the user, and for the training set, I will use the dictionary of metrics computed from the Polígrafo and Fake.br.Corpus, having the emotion metrics from the Polígrafo dataset and the remaining from Fake.br.Corpus.

[8]https://poligrafo.sapo.pt/

[9]https://matplotlib.org/

[10]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

	Precision		Recall		F1 Score		Accuracy
Indicators categories	False	True	False	True	False	True	
Emotion	0.67	0.67	0.677	0.67	0.67	0.67	0.67
Subjectivity	0.40	0.00	0.67	0.00	0.50	0.00	0.33
Affectivity	1.00	0.75	0.67	1.00	0.80	0.86	0.83
Polarity	0.00	0.40	0.00	0.67	0.00	0.50	0.33
BP	0.25	0.00	0.33	0.00	0.29	0.00	0.17
Grammatical	0.00	0.33	0.00	1.00	0.00	0.50	0.33
Combined Indicator	1.00	0.60	0.33	1.00	0.50	0.75	**0.67**

(a) Indicators performance in Polígrafo.

	Precision		Recall		F1 Score		Accuracy
Indicators categories	False	True	False	True	False	True	
Emotion	0.63	0.59	0.48	0.72	0.54	0.65	0.60
Subjectivity	0.53	0.55	0.55	0.52	0.54	0.54	0.54
Affectivity	0.93	0.93	0.92	0.94	0.93	0.93	0.93
Polarity	0.74	0.77	0.78	0.74	0.76	0.75	0.76
BP	0.65	0.65	0.63	0.68	0.64	0.66	0.65
Grammatical	0.88	0.92	0.93	0.88	0.90	0.90	0.90
Combined Indicator	0.96	0.96	0.96	0.96	0.96	0.96	**0.96**

(b) Indicators performance in Fake.br.Corpus.

Table 1. Indicators performance in forecasting news veracity using Logistic Regression.

Logistic Regression also allows us to predict the probability of an event's occurrence; in this case, the probability of an article being false or true. This allows us to formulate the disinformation detection task in FactMe as a classification task. As the Logistic Regression can model the probability of a particular class, I will estimate the likelihood of an article being false for each linguistic indicator and call it as **indicator score**. I used the dictionary of metrics computed from the articles datasets for the training set and created a classification model for each indicator category. Each classification model will be trained with the set of features respective to the indicators' category taken from the computed dictionary of metrics. The emotion metrics will be taken from the Polígrafo dataset, and the remaining will be taken from Fake.br Corpus. To predict the probability of each class (true or false), I will use predict_prob, a Logistic Regression function from sklearn that gives the probability for the target (true or false in our case) in an array form. The predict_prob function will be used on each model and will receive as input each indicator's metrics computed from the news article given by the user to assess in FactMe. The result of predict_prob on each model will be the probability of the article being false, thus presenting each indicator score to the user. Regarding the combined indicator, a model will be created using as features, all the metrics gathered from the dictionary of metrics computed from the Fake.br Corpus (as tested in *Performance of linguistic indicators* section). It will also be used predict_prob function receiving as input, all the metrics computed from the news article given by the user to assess in FactMe.

Computation of disinformation score

To help users judge the credibility of the article better, FactMe will also present a disinformation score in a progress bar in a color gradient scale: green, yellow, and red as presented in *FactMe user interface* section. The color will rank an article from most credible (green) to less credible (red). The colors are determined by the probability of the article being false with less or equal than 33% belonging to green, between 33% and 66% belonging to yellow, and 66% or more belonging to red. The probability will be given by a weighted sum of the combined indicator and the source credibility indicator. I will give the combined indicator a weight of 75% and the source credibility indicator a weight of 25%. This means that the probability of the combined indicator calculated previously will be considered 75%, and it will increase the falsehood probability (disinformation score) by 25% if the article is not trusted. In such manner, I will **rank an article** and present the disinformation score to the users.

Concerning the **implementation** of FactMe, the front-end is implemented with Flask (web framework written in Python), HTML and Javascript. For the back-end, I used Python Programming Language.

FactMe is hosted in a server provided by INCD in the following link: `http://194.210.120.9/`.

EVALUATION

Evaluation of the disinformation indicators

In this evaluation, the user has to give an opinion about an article that he reads. Six news articles were taken from Polígrafo, an online journalistic project that does fact-checking to Portuguese news articles. I took three news articles classified as "True" or "True, but" and three news articles classified as "False" or "Pepper on the tongue" in the rating scale of Polígrafo, and I classified them as true articles and false articles, respectively, and used them for the questionnaire.
In this questionnaire, the A/B testing approach was implemented, consisting of a randomized experiment with two variants, A and B. With this approach, the subject's answers to variant A and variant B can be compared to determine which

of the two is more effective. Variant A corresponds to a questionnaire about the presented article without indicators, and variant B corresponds to a questionnaire with indicators.

When we make the user answer the presented article's questions without any further information (variant A), an opinion about the article is created by the user, which is again questioned in the B variant when presented with the indicators. The method is different from presenting the article with the indicators (variant B) in the first place because it creates an opinion about the article immediately considering the indicators. This observation made me create the following 2 versions of questionnaires to be answered alternatively in order to compare them later:

1. Version 1 shows a variant A which presents an article to the user without any indicators and a set of questions related to the article; after answering that questions, users are presented with B, the second variant where it shows the same article but this time with the computed indicators.

2. Version 2 directly presents the articles to the user with the indicators also shown as variant B of the first version.

We made a user answer only one of these versions, and it is assured that we have the same amount of answers for both versions to compare later.

For the first version of the assessment, the questionnaire is divided into A and B tests. In A, the user has to read the news article and answer the questions related to the article. In B, the same article is presented along with the indicators and scores computed by the tool. The user is asked the same questions from variant A and additional questions about the indicators' impact. The second version of the assessment presents the B variant of the first version directly.

There are in total six questionnaires, each one having a different news article. The user answers the questionnaire one at a time (one of the versions). At the end of the questionnaire, the user is asked if he wants to answer one more, showing the questionnaires left to answer. The evaluation page uses session cookies to record filling the id of the questionnaire answered by the user. The cookies will allow users to continue the questionnaire where they left it in the last session.

With this evaluation and regarding the two versions that we have, two types of observations can be taken into consideration:

1. The impact can be measured by showing the same news without indicators and then with indicators (comparison between A and B).

2. The impact can also be assessed by measuring the difference in presenting news with the first and second versions.

This assessment aims to evaluate the impact of the developed indicators on the users by presenting them the alternative approaches.

Results

I have conducted an evaluation where six articles from Polígrafo were evaluated, three articles with true veracity and three with false veracity. This evaluation was made by 16 people, obtaining 34 articles evaluated for version 1 for each variant and 33 for version 2. Table 2 shows the results of both versions for the eight questions.

Regarding the **first question**, The answers for the true articles presented without indicators are mostly somewhat and completely representative but, after showing the indicators, the answers spread by the other options decreasing the representation of the title. In the false articles, variant A shows that users saw the title as mostly representative, and in variant B of version 1, the completely representative assessment increased along with the completely unrepresentative. In version 2, the false article also showed a more representative title than the true articles. Overall, the title's representation was not interpreted by the users according to the articles' classification, and the indicators did not make a big difference.

The **second question** is correlated with the first one. As we saw in the first question, a large majority answered somewhat or completely representative. The second question's correlation with the first question is reflected in the answers, with most answers being A (it is representative) in both versions. Options C (Title carries only a little information about the body) and E (Title overstates/understates claims or conclusions) were also selected from a few respondents. The true articles had mostly a representative title but as well as the false one.

The **third question** has as options 1 to 5 where 1 corresponds to not clickbaity, and 5 corresponds to very clickbaity. Variant A shows that the true articles are not that clickbaity being the majority of the responses divided between 1 & 3. The false articles had a big part of the responses in 4 & 5. After showing the indicators, no significant changes were seen; we noticed a small rise in false articles in option 5 and also a small rise in option 4 for the true articles. In version 2, the values are similar, being distributed by the options highlighting the true articles as not clickbaity and false articles as very clickbaity. Overall the indicators identified some false articles as clickbaity but also generated some doubts about the true ones.

The **fourth question** (If you think the title is clickbaity, what makes you think it is?) was an open question. It was observed that some respondents found the title of the true articles clickbaity because the names used in the article's title were known. It can also be noticed that the false articles' title captures much more attention. Overall, the number of answers about the clickbait was much higher in the false articles than in the true, showing that the respondents doubted false articles and identified some as more clickbaity.

The **fifth question** has as options 1 to 5 where 1 corresponds to not subjective, and 5 corresponds to very subjective. It was observed that without indicators, the false articles tend to be more subjective and true articles less subjective, being the true ones distributed mostly in 1 and 3. After showing the indicators, the false articles' distribution remained more or less unchanged, but the true articles got higher scores from the respondents, being more subjective. For version 2, we

Questions	Version 1 - A		Version 1 - B		Version 2 - B	
1 - Does the title represent the content of the article?	True	False	True	False	True	False
C. Unrepresentative	2	0	4	2	2	1
S. Unrepresentative	3	5	4	4	4	1
S. Representative	6	5	5	2	5	5
C. Representative	6	7	4	9	6	9
2 - Why the title does not represent the content of the article?						
It is representative	11	11	9	11	8	11
Title is on a different topic than the body	0	0	2	0	0	0
Title carries only a little information about the body	3	3	3	2	5	0
Title takes a different point of view than the body	1	0	1	0	1	2
Title overstates/understates claims or conclusions	2	3	2	4	3	3
3 - Is the title clickbaity?						
1	6	2	6	2	6	2
2	4	1	2	1	0	2
3	4	3	4	4	5	3
4	1	6	3	3	3	3
5	2	5	2	7	3	6
5 - Is the article subjective?						
1	6	2	3	1	7	5
2	1	1	3	2	4	0
3	6	3	4	4	4	2
4	1	4	2	2	0	4
5	3	7	5	8	2	5
6 - How do you classify this news?						
True	7	3	6	0	7	2
Dubious	9	8	9	6	10	5
False	1	6	2	11	0	9
7 - Did the indicators made an impact in the credibility of the article?						
Big impact	-	-	8	7	8	10
Small impact	-	-	6	8	9	4
No impact	-	-	3	2	0	2
8 - Which indicator made the biggest impact?						
Emotion	-	-	0	2	2	2
Subjectivity	-	-	4	0	5	0
Affectivity	-	-	0	0	0	2
Polarity	-	-	2	0	1	0
BP	-	-	0	0	1	1
Source	-	-	8	13	8	11
None	-	-	3	2	0	0

Table 2. Results of the evaluation of disinformation indicators.

noticed an improvement towards the true articles being divided between options 1 and 3, proving to be less subjective. The majority of the false articles were also distributed from 3 upwards. In general, the evaluation without indicators (variant A) got good results, and the variant B of version 2 had a better subjectivity distribution than version 1.

Regarding the **sixth question**, in variant A of version 1, the false articles were mostly classified as dubious and false; however, three false articles were considered true. The true articles are distributed between true and dubious labels. We can see that more false articles were classified as false after showing the indicators, and few more true articles classified as dubious. In version 2, we can see an improvement in the true articles without being classified as false, although some false articles were classified as true. Overall, we can see that more false articles were classified correctly as false in both versions with indicators, but version 2 classified better the true articles, not placing the false classification wrongly.

The **seventh question** concerns the impact of the indicators. We can see that most of the respondents classified the impact as small and big for both versions. Version 1 had some answers stating that the indicators did not influence their perceptions of the article's credibility. Version 2 had a more significant impact on false articles and a little more influence on true articles. Overall, the results show us that the indicators influenced the users in evaluating true and false articles.

Regarding the **eighth question**, in both versions, we can see that the source credibility indicator had the most significant impact on the users for both types of articles. Apart from that, other indicators also had some influence, namely subjectivity. In version 1, there were some assessments where none option was given for both labels, unlike Version 2. This could lead us to think that some users may not have changed their opinion about the article's veracity after visualizing the indicators.

Overall, we can see that showing the indicators did not influence users on their assessment of the news title's representation, but it helped to classify the articles. Showing the indicators in both versions made a difference, but version 2 had slightly better results for the articles' classification than version 1 variant B. It can be concluded that showing the indicators made at least a small impact on the articles' classification. Finally, the source credibility indicator had the most significant impact in judging articles' credibility.

Evaluation of the user interface

Questionnaire

To evaluate the user interface, I used the QUIS (Questionnaire For User Interaction Satisfaction). QUIS was originated from a team of researchers in the HCIL at the University of Maryland and is currently in version 7.0 [4]. QUIS measures the following six aspects: (1) Reaction to software, (2) Screen, (3) Terminology and system information, (4) Learning, (5) System capabilities, (6) Usability & user interface.

One aspect measures the overall satisfaction called reaction to software, and the other five aspects measure five dimensions of the interface. For each aspect, there is a section that measures factors regarding that specific dimension of the interface. The aspects are measured with a 9-points rating scale. The lowest and the highest values are associated with a word to describe the user assessment. The questionnaire also includes two open questions regarding the positive and negative aspects of the system under evaluation.

Aspects	Average	Median
Software reaction	8.73	9
Screen	8.55	9
Terminology and system information	8.40	8
Learning	8.43	9
System capabilities	8.13	8
Usability and user interface	8.61	9
Global average	8.48	9

Table 3. Results of the evaluation of user interface with QUIS.

It was made in Google Forms, and I also included a link to the form in FactMe in the "evaluation" section. The form was created in the Google Cloud of Instituto Superior Técnico, and the assessment was anonymous; no personal information was captured from the user.

Results

Table 3 presents the summarized results of the QUIS with the average and median of each aspect and a global average. The assessment was made by 11 people. Is was calculated the average and median of each question and each aspect, and a global average. Overall, the users evaluated the interface of FactMe positively with an average of 8.48.

The results show that the reaction to software aspect was good in general, having an average above 8 in every rating.

Regarding the screen components, the items are considered easy to find, the information is well organized, and the font and size are adequate. The highlighting on the screen has the lowest rating among other factors, leaving room for improvement in this factor.

The terminology and system information factors were evaluated positively, having an average above 8 in every rating and a median of 8.

The learning factors had mostly positive ratings, but the help messages had the lowest rating (7.73), needing to be improved in the tool.

The system capabilities also had an average above 8, but it was slightly below the other groups. A few users considered that the users should have a little experience. This factor can be improved by adding more help messages, which was also one of the improvements to FactMe.

Concerning the usability and interface factors, the users agreed to have adequate use of colors, a good response to errors, and system messages giving, in general, good feedback to the system.

Additionally, the questionnaire also included two questions concerning the negative and positive aspects. Users referred to simplicity, effective, well-articulated, and organized system. They also mentioned that the tool is easy to use and navigate.

Some improvements were also noted, including a more flashy homepage, a homepage button which could be better noticed and, a more responsive smartphone design.

CONCLUSIONS

In this work, I have surveyed previous research on content-based (linguistic) and context-based metrics, including some disinformation detection models, and have also reviewed similar tools to FactMe. This research helped me compute the linguistic metrics and the context indicators for FactMe. I created a dataset of news articles from Polígrafo and also used the Fake.br Corpus to create a dictionary with the computed metrics for the articles in both datasets. Using the dictionary, I calibrated the linguistic indicators and developed a learning model using Logistic Regression. The model led us to achieve an accuracy of 96% with the combined indicator in predicting the article's veracity. Related tweets and source credibility were also computed as context-based indicators.

Regarding the user interface of FactMe, it presents to the users the score of each linguistic indicator in a color gradient scale, the related tweets, and source credibility indicator. Additionally, a disinformation score will also be computed, using the combined indicator and the source credibility to present to consumers also in a color gradient scale.

Finally, I made two evaluations of FactMe regarding the indicators' impact and the user interface using QUIS. I can say that FactMe is a simple and easy visualization tool that made, for some users at least, a small difference in judging the articles and being aware of disinformation consumption.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (May 2017), 211–36. DOI: http://dx.doi.org/10.1257/jep.31.2.211

[2] Danielle Caled and Mário J. Silva. 2020. *Linguistic disinformation metrics for Portuguese*. Technical Report.

[3] Paula Carvalho, Bruno Martins, Hugo Rosa, Silvio Amir, Jorge Baptista, and Mário J. Silva. 2020. Situational Irony in Farcical News Headlines. In *Computational Processing of the Portuguese Language*. Springer International Publishing, 65–75.

[4] John P. Chin, Virginia A. Diehl, and Kent L. Norman. 1988. Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '88)*. Association for Computing Machinery, New York, NY, USA, 213–218. DOI:http://dx.doi.org/10.1145/57167.57203

[5] P. Ekman. 1992. An argument for basic emotions. *Cognition Emotion* 6 (1992), 169–200.

[6] Don Fallis. 2015. What Is Disinformation? *Library Trends* 63 (01 2015), 401–426. DOI: http://dx.doi.org/10.1353/lib.2015.0014

[7] Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Jarvelin, Rosie Jones, YiquN Liu, Josiane Mothe, Wolfgang Nejdl, Isabella Peters, and Benno Stein. 2018. An Information Nutritional Label for Online Documents. *SIGIR Forum* 51, 3 (Feb. 2018), 46–66. DOI:http://dx.doi.org/10.1145/3190580.3190588

[8] B. Liu. 2010. *Sentiment analysis and subjectivity*. 627–666.

[9] Rafael Monteiro, Roney Santos, Thiago Pardo, Tiago Almeida, Evandro Ruiz, and Oto Vale. 2018. *Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*. 324–334. DOI:http://dx.doi.org/10.1007/978-3-319-99722-3_33

[10] Charles Osgood, George J Suci, and Percy H Tannenbaum. 1957. The measurement of meaning. *University of Illinois Press* (1957).

[11] Ellen Riloff and Janyce Wiebe. 2003. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP '03)*. Association for Computational Linguistics, USA, 105–112. DOI: http://dx.doi.org/10.3115/1119355.1119369

[12] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19 (08 2017), 22–36. DOI: http://dx.doi.org/10.1145/3137597.3137600

[13] Lina Zhou, Judee Burgoon, Douglas Twitchell, Tiantian Qin, and Jay Jr. 2004. A Comparison of Classification Methods for Predicting Deception in Computer-Mediated Communication. *J. of Management Information Systems* 20 (03 2004), 139–165. DOI: http://dx.doi.org/10.1080/07421222.2004.11045779

[14] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and Resolution of Rumours in Social Media: A Survey. 51, 2, Article 32 (Feb. 2018), 36 pages. DOI: http://dx.doi.org/10.1145/3161603