Extraction and Visualization of Fake News Indicators

Karan Prem Manghnani

Thesis to obtain the Master of Science Degree in
Information Systems and Computer Engineering

Supervisors: Prof. Mário Jorge Costa Gaspar da Silva
Danielle Cañed Vieira

Examination Committee

Chairperson: Prof. José Carlos Martins Delgado
Supervisor: Prof. Mário Jorge Costa Gaspar da Silva
Member of the Committee: Prof. Bruno Emanuel Da Graça Martins

December 2020
Acknowledgments

First of all, I would like to thank my supervisor Prof. Mário Gaspar da Silva for the countless hours of meetings, support and feedback. I could not have imagined having a better advisor and mentor. I am also thankful to my supervisor Prof. Danielle Caled for the continuous support and guidance. Their knowledge and insight were crucial to the development of this thesis.

I am very thankful to my uncle Francisco Barbosa for being my mentor, supporting me and making me capable to complete this journey.

I would like to thank INCD (Infraestrutura Nacional de Computação Distribuída) for providing me a host for this project.

I would like to thank all my family, long date friends from primary & high school and colleagues for supporting me and contributing to this thesis in a way or another.

A special thanks to my friends Rodrigo Nobre and Fábio Pimenta for motivating me, supporting me, for the sleepless nights we were working together before deadlines, and for all the fun we had in the last five years.

Finally and above all, I am forever thankful to my parents for their unconditional support and love, not only during my academic journey but during all my life.

To each and every one of you – Thank you.
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.

Keywords

Disinformation indicators; Disinformation visualization; Fake news.
Resumo

Esta dissertação apresenta FactMe, uma ferramenta de visualização online que ao receber uma notícia, calcula uma série de métricas destinadas a fornecer aos consumidores, indicadores de qualidade da informação. O consumo de notícias na Internet está a aumentar rapidamente e torna-se cada vez mais difícil avaliar a confiabilidade das notícias encontradas na web. Como a web é aberta, essa vantagem pode contribuir para a propagação da desinformação. Nos últimos anos, a desinformação tem causado graves danos e pode causar ainda mais futuramente. Isso mostra o quão vital pode ser uma ferramenta de detecção de desinformação. Avanços recentes permitem-nos analisar notícias e usar processamento de linguagem natural e algoritmos de inteligência artificial para detetar desinformação e ajudar os consumidores a avaliar a credibilidade dos artigos. Neste trabalho, fiz um levantamento de pesquisas anteriores sobre métricas baseadas no conteúdo (linguísticas) e baseadas no contexto, o que me ajudou a desenvolver um modelo de aprendizagem usando os indicadores linguísticos que alcançaram uma exatidão de 96% na previsão da veracidade dos artigos. Além disso, este trabalho fornece uma avaliação na perspetiva do consumidor, que mostrou uma satisfação do usuário positiva com uma pontuação média de 8,48 no QUIS e pelo menos um pequeno impacto dos indicadores de desinformação sobre os consumidores na previsão da veracidade do artigo.

Palavras Chave

Indicadores de desinformação; Visualização de desinformação; Notícias falsas.
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.6 Summary</td>
<td>40</td>
</tr>
<tr>
<td>4 Evaluation</td>
<td>43</td>
</tr>
<tr>
<td>4.1 Evaluation of the disinformation indicators</td>
<td>45</td>
</tr>
<tr>
<td>4.1.1 Characterization of users</td>
<td>45</td>
</tr>
<tr>
<td>4.1.2 Questionnaire</td>
<td>45</td>
</tr>
<tr>
<td>4.1.3 Results</td>
<td>47</td>
</tr>
<tr>
<td>4.2 Evaluation of the user interface</td>
<td>52</td>
</tr>
<tr>
<td>4.2.1 Questionnaire</td>
<td>52</td>
</tr>
<tr>
<td>4.2.2 Results</td>
<td>52</td>
</tr>
<tr>
<td>4.3 Summary</td>
<td>55</td>
</tr>
<tr>
<td>5 Conclusion</td>
<td>57</td>
</tr>
<tr>
<td>5.1 Contributions</td>
<td>59</td>
</tr>
<tr>
<td>5.2 Future work</td>
<td>59</td>
</tr>
<tr>
<td>A Evaluation questionnaire of indicators</td>
<td>65</td>
</tr>
<tr>
<td>B Evaluation questionnaire of user interface (QUIS)</td>
<td>69</td>
</tr>
<tr>
<td>C Prototypes</td>
<td>75</td>
</tr>
<tr>
<td>D FactMe Navigational components interface</td>
<td>79</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Mockup of an initial user interface for FactMe. ................................................. 5
2.1 Twitter Verification Assistant framework. .......................................................... 16
2.2 Interface of Twitter Verification Assistant. ......................................................... 17
2.3 Architecture of XFake. ...................................................................................... 18
2.4 XFake’s interface for prediction and explanation. .............................................. 18
2.5 Screenshot of the NewsGuard extension interface. ............................................. 19
2.6 Screenshot of the FactMata interface. ................................................................. 20
3.1 Information processing of FactMe. .................................................................... 25
3.2 FactMe interface. ............................................................................................... 27
3.3 Boxplot of emotion metrics. .............................................................................. 32
3.4 Boxplot of subjectivity metrics. ......................................................................... 33
3.5 Boxplot of affectivity metrics for Poligrafo and Fake.br Corpus. ....................... 34
3.6 Boxplot of polarity metrics. .............................................................................. 35
3.7 Boxplot of BP metrics. ...................................................................................... 36
3.8 Boxplot of grammatical metrics. ....................................................................... 36
3.9 Computation of indicator scores and the final disinformation score in FactMe. ... 39
4.1 Alternative questionnaires for the evaluation of the FactMe disinformation indicators. ... 46
4.2 Distribution of the answers related to the title of the article. ............................. 48
4.3 Distribution of the answers related to subjectivity and classification of the article. ... 50
4.4 Distribution of the answers related to the impact of the indicators. ..................... 51
A.1 Evaluation without indicators questionnaire (1). .............................................. 65
A.2 Evaluation without indicators questionnaire (2). .............................................. 66
A.3 Evaluation without indicators questionnaire (3). .............................................. 66
A.4 Evaluation with indicators questionnaire (1). ................................................. 67
List of Tables

3.1 Indicators performance in forecasting news veracity using Logistic Regression. . . . . 38
4.1 Questionnaire of variant A and variant B for evaluation of disinformation indicators. . . . 47
4.2 Answers to the question "What makes you think the title is clickbaity?". . . . . . . . . . . 49
4.3 Results of FactMe evaluation with QUIS. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 54
# Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>URL</td>
<td>Uniform Resource Locator</td>
</tr>
<tr>
<td>BP</td>
<td>Behavioral and Physiological</td>
</tr>
<tr>
<td>PoS</td>
<td>Part-of-Speech</td>
</tr>
<tr>
<td>NER</td>
<td>Named-entity recognition</td>
</tr>
<tr>
<td>LIWC</td>
<td>Linguistic Inquiry and Word Count</td>
</tr>
<tr>
<td>SVM</td>
<td>Support-Vector Machines</td>
</tr>
<tr>
<td>BoW</td>
<td>Bag-of-words</td>
</tr>
<tr>
<td>NLTK</td>
<td>Natural Language Toolkit</td>
</tr>
<tr>
<td>HITS</td>
<td>Hyperlink-Induced Topic Search</td>
</tr>
<tr>
<td>TEXTA</td>
<td>Terminology Extraction and Text Analytics</td>
</tr>
<tr>
<td>TB</td>
<td>Tweet-based</td>
</tr>
<tr>
<td>UB</td>
<td>User-based</td>
</tr>
<tr>
<td>ID</td>
<td>Identifier</td>
</tr>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>ANEW</td>
<td>Affective Norms for English Words</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>MPQA</td>
<td>Multi-Perspective Question Answering</td>
</tr>
<tr>
<td>ERC</td>
<td>Entidade Reguladora para Comunicação Social</td>
</tr>
<tr>
<td>INCD</td>
<td>Infraestrutura Nacional de Computação Distribuída</td>
</tr>
<tr>
<td>WSGI</td>
<td>Web Server Gateway Interface</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>HTML</td>
<td>Hypertext Markup Language</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>HCIL</td>
<td>Human–Computer Interaction Lab</td>
</tr>
<tr>
<td>JSON</td>
<td>JavaScript Object Notation</td>
</tr>
<tr>
<td>QUIS</td>
<td>Questionnaire for User Interface Satisfaction</td>
</tr>
<tr>
<td>COVID-19</td>
<td>Coronavirus Disease 2019</td>
</tr>
</tbody>
</table>
Introduction

Contents

1.1 Objectives ................................................. 4
1.2 Methodology ............................................. 4
1.3 Thesis outline .......................................... 5
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.

According to Zubiaga et al. (2018), social media has become a critical publishing tool for journalists and the primary consumption method for citizens looking for the latest news. 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 (Fallis, 2015). The terms misinformation and disinformation have often been associated with the term fake news. According to Allcott and Gentzkow (2017), fake news is "a news article that is intentionally and verifiably false". 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 BBC1, 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, by Shu et al. (2017). 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 like FactMata\(^2\) and NewsGuard\(^3\) have been made using many different approaches to help readers detect and judge online news by presenting indicators.

### 1.1 Objectives

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.

### 1.2 Methodology

This accomplished work has taken the following steps:

1. Study previous research on content-based and context-based disinformation metrics, including models for fake news classification.

2. Review similar tools to understand their organization, computed metrics, and user interface.

3. Establishment of the metrics to be used.

4. Design of FactMe architecture.

5. Development of a digital and paper prototype of the intended interface of disinformation detection of FactMe. (See on Figure 1.1 an initial mockup of the user interface. The remaining prototypes can be found in appendix C).

6. Collection of news articles from fact-checking site to create a dataset.

7. Initial implementation of the metrics.

8. Experimental study of the metrics for classification of news articles in different datasets using boxplots to display distribution and machine learning algorithm.

---

\(^2\)https://factmata.com/

\(^3\)https://www.newsguardtech.com/

10. Evaluation of FactMe under a consumer perspective.

1.3 Thesis outline

The dissertation is organized into five chapters. Chapter 2 reviews previous research in content-based and context-based metrics and surveys the user interfaces and functionality of similar tools. Chapter 3 presents the information processing of FactMe, the user interface, the indicators and their computation, a calibration of the metrics and indicators, and the implementation of FactMe. Chapter 4 presents the results obtained by the user assessment and concludes about the usefulness of the tool. Finally, Chapter 5 overviews the main conclusions and contributions regarding this project and proposes future work.
## Related Work

### Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Linguistic disinformation metrics and indicators</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Complex indicators</td>
<td>12</td>
</tr>
<tr>
<td>2.3 Examples of related software</td>
<td>14</td>
</tr>
<tr>
<td>2.4 Summary</td>
<td>21</td>
</tr>
</tbody>
</table>
This chapter overviews relevant related work that helped construct FactMe, including disinformation metrics, indicators, and similar tools for detecting disinformation.

2.1 Linguistic disinformation metrics and indicators

Linguistic metrics can be grouped into six categories of indicators described below:

**Affectivity Indicator:** Disinformation often employs a higher number of emotional words than informative text (Fuhr et al., 2018). 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. (1957) 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 (Zhou et al., 2004). The BP category corresponds to six processes such as:

1. **Biological:** e.g. body, health, sexual;
2. **Cognitive:** e.g. insight, causation, discrepancy, certainty, inclusion and exclusion;
3. **Perceptual:** e.g. sounds, smells, physical sensations;
4. **Personal:** e.g. work, achievement, leisure, money, religion;
5. **Relative:** e.g. motion, spatio-temporal locations of people;
6. **Social:** e.g. family, friends;
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.

**Emotion indicator:** One characteristic of disinformation is that it may make an inflammatory emotional appeal to the reader (Fuhr et al., 2018). 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 (1992):

1. Anger;
2. Disgust;
3. Fear;
4. Happiness;
5. Sadness;
6. 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 Part-of-Speech (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. (2004) and Carvalho et al. (2020) 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 (degree of emotiveness in a text): \( \frac{\text{sum of occurrences of adjectives and adverbs}}{\text{sum of occurrences of nouns and verbs}} \)
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 (frequency of pauses in a text): \( \frac{\text{number of punctuation signals}}{\text{total number of sentences}} \)
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 (Liu, 2010).

Carvalho et al. (2020) proposed the following metrics for this category:

1. Polarity information: \[ \frac{\text{number positive and negative words}}{\text{number of words}} \]

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. Riloff and Wiebe (2003) stated that 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) and proposed the following metrics:

1. Subjectivity ratio: \[ \frac{\text{strong or weak subjective words}}{\text{number of words in the text}} \]

2.1.1 Evaluation of metrics for disinformation detection

To evaluate the feasibility of the linguistic metrics in disinformation detection, Caled and Silva (2020) made an experimental study using the Fake.Br Corpus (Monteiro et al., 2018), which contains news articles written in Brazilian Portuguese (3600 True + 3600 False) from 2016 to 2018 that were manually labeled news. They used the SVM classifier (the LinearSVC implementation in Scikit-learn, with default parameters), reporting the general accuracy, precision, recall, and F1 score metrics for both classes and a partition of 0.8/0.2 in testing/training sets. The following approaches were tested:

1. Baseline: BoW approach made by Monteiro et al. (2018) with the extraction of binary occurrences of unigrams, stopword, punctuation removal and stemming.

2. Simple BoW: Removed the punctuation and counted the occurrence of each unigram token in the document.

3. All Linguistic Metrics.

4. All Linguistic Metrics + Baseline.

5. All Linguistic Metrics + Simple BoW.

The authors used different settings of metrics evaluated on the Fake.br Corpus. The obtained results showed that using only the baseline metrics could not lead to good results. However, using the simple BoW approach, the results improved slightly. The best approach was obtained using the linguistic metrics; it gave better accuracy, precision, recall, and F1 score results.
Regarding the linguistic metrics, the authors also conducted an ablation study by removing each linguistic metric while keeping the remaining ones. They noticed that the individual removal of each metric set had a minimal impact on accuracy. The authors also removed certain sets of metrics choosing the linguistic metrics sets that yield the best results, i.e., affectivity, grammatical, sentiment polarity.

The authors observed that removing the best sets of metrics (affectivity and grammatical metrics) and (affectivity, grammatical, and sentiment polarity metrics) decreased the accuracy.

### 2.2 Complex indicators

This section presents higher-level indicators proposed by Fuhr et al. (2018) as Information Nutrition Label, resembling nutrition fact labels on food packages. Such a label describes, along with a range of agreed-upon dimensions, the contents of the product (an information object, in our case) to help the consumer (reader) decide about the consumption of the object. As information nutrition labels, the authors proposed some of the following categories presenting a brief description and a method to implement the indicator in English:

**Factuality:** This indicator concerns the factuality of the sentences from the author’s perspective and not its truthness. Factual statements are different from opinions; they are written in a certain way by the author that can be detected. Terms like shall, can and will indicate the author’s confidence, which can evaluate the factuality. For the factuality prediction, there are two approaches:

- Rule-based: Uses TruthTeller algorithm (Lotan et al., 2013) to assign a score of factuality to propositions.

- Machine learning-based: Lee et al. (2015) used a SVM regression model to predict continuous factuality values from shallow lexical and syntactic features such as lemmas, PoS tags, and dependency paths.

Both approaches can be combined; the Truhtteller algorithm’s output is used as features for the SVM regression model to predict the final factuality value. For factuality inference, Fuhr et al. (2018) also suggest the use of datasets such as Fact-Bank (Saurí and Pustejovsky, 2009) and MEANTIME (Minard et al., 2016), which can capture the factuality statements.

**Readability:** The goal of this indicator is to estimate how easily a reader can understand a written text. It depends on the following three categories:

- Writing quality: Refers to the grammatical correctness of the text.

- Targeted audience: Refers to the educational background required to understand the text.

- Presentation: Refers to typographic aspects like font size, line height, and line length.
The current measures to understand the written text are linear regression based on the number of words, syllabus, and sentences (Collins-Thompson, 2014). Coh-Metrix is a computational tool that measures cohesion and text difficulty at various language, discourse, and conceptual analysis levels, and it was developed by Crossley et al. (2008).

**Virality:** News can be spread virally from a dubious source across the internet. This indicator analyzes the information flow and information objects across the internet. For the virality detection on texts and information, there are four types of property sets to take into account:

- The sender.
- The information object.
- The recipient.
- The channel in which the information is spread.

Combining these sets will help determine virality, decide how fast the information will flow and how far it can reach. One of the platforms that help news go viral is social media. There are many examples of signals that can detect virality in social media, e.g., comments, the number of likes, retweets speed. There are datasets of tweets Arturo Deza (2020) and viral images Weng et al. (2014) that can be used for training and also software¹ (e.g., Hoaxy, a software proposed by Indiana University) to visualize the spread of claims.

**Opinion:** Is an element of the text which reflects the author’s opinion. An author may push a particular viewpoint that is not explicitly expressed in the text. In this category the task is to detect the sentences that are opinionated and it can be calculated in the following way:

\[
\text{Opinion ratio} = \frac{\text{Opinionated sentences}}{\text{Total sentences}}
\]

In order to get the opinionated sentences, online software can be used e.g. NLTK opinion lexicon reader² or Opinion Finder³.

**Authority / Credibility / Trust:** Trust and authority, as synonyms, in this case, refer to a property of the source of a message, while credibility is an attribute of the message itself. When a user analyzes a news article, they are mainly interested in the credibility of the news. To judge credibility, they usually check the trustworthiness of the source, and it is easier to estimate the authority of the source than credibility. The goal of this indicator will be to determine the authority or trust of the source. The sources can be websites on which the authority can be estimated using PageRank or HITS algorithm and social media tools like Tweetcred⁴ that can estimate credibility.

¹https://hoaxy.iuni.iu.edu/
²http://www.nltk.org/_modules/nltk/corpus/reader/opinion_lexicon.html
³https://mpqa.cs.pitt.edu/opinionfinder/
⁴https://chrome.google.com/webstore/detail/tweetcred/fboklijnojeihdnkieeneiakdgikg
**Technicality:** The content of the article can have technical terms and only be understandable to certain people who have knowledge in a particular field. The goal of this indicator is to calculate the technicality and understand how hard it should be to understand the news article. Like readability, but more focused on the content than form, it will capture the proportion of the domain specific vocabulary used by the document. For that, there is available software to extract technical words for technicality measurement, for example, TEXTA (Terminology extraction and text analytics).

**Topicality:** Topical documents are documents that cover topics that are in the current zeitgeist (trend), so this will decide if the document is of current interest or not. Fake news is often falsely influenced by the current news cycle. One of the methods can be to extract the key terms and entities of the news articles and compare that to recent news or publications by searching on social media or search engines, knowing that these indicators’ results vary from time to time. For that purpose it can be used online text mining tools and key phrase extraction.

### 2.3 Examples of related software

This section reviews examples of some software tools made to detect disinformation.

#### 2.3.1 Twitter Verification Assistant

Popular social media platforms such as Twitter and Facebook are now an integral part of the journalistic and news diffusion process, not only because these platforms have lowered the barrier for citizens to contribute to news generation and documentation with their own content, but also due to the possibilities they offer for rapidly disseminating news to one’s network of contacts and broader communities. These new publishing capabilities and sharing content have led to an uncontrolled propagation of large volumes of news content over social networks quickly, creating a big challenge for verification’s journalistic process. Having in mind these challenges, Boididou et al. (2017) presented a software called Twitter Verification Assistant that supports the automatic classification of multimedia Twitter posts as credible or misleading. The overview of the framework can be seen in Figure 2.1.

A work by Gupta et al. (2014) showed a system trained on a set of features to discriminate between tweets sharing fake images and tweets sharing real images on a dataset of tweets related to Hurricane Sandy. In that way, tweet classification was used as a first step towards verifying the associated images. However, the separation between training and test cases was not adequate for reliably assessing the method’s generalization ability. This led to Boididou et al. (2017) move one step forward and present

---

5[https://github.com/texta-tk/texta](https://github.com/texta-tk/texta)
7[https://github.com/snkim/AutomaticKeyphraseExtraction](https://github.com/snkim/AutomaticKeyphraseExtraction)
8[http://reveal-mklab.iti.gr/reveal/fake/](http://reveal-mklab.iti.gr/reveal/fake/)
an approach that uses a variety of content-based and context-based features for the social media post
in question and builds two classification models used to produce two independent first-level predictions
regarding the credibility of the post. At a second step, a top-level classifier leverages these first-level pre-
dictions on "unseen" content for retraining the best of the first-level models, following a semi-supervised
learning paradigm. In this way, the system produces more confident predictions. The two indepen-
dent classification models use two different set of features: tweet-based (TB) and user-based (UB). Tweet-based features consider four types of features:

1. **Text-based**: extracted from the text, it considers the length, number of words, questions, excla-
mations;
2. **Language-specific**: includes the number of positive and negative sentiment words, nouns, slang
   words and also personal pronouns;
3. **Twitter-specific**: includes retweets, mentions, hashtags;
4. **Link-based**: provides information from links shared by tweets.

User-based features consider two types of features:

1. **User-specific**: includes the user’s number of friends and followers, account age, number of tweets,
   and if it is a verified account by Twitter;
2. **Link-based**: depending on the existence of a URL in the Twitter, it is applied the same Web metrics
   like the ones used in the link-based of TB features.

Boididou et al. (2017) use TB and UB features to build two independent classifiers (CL1, CL2, re-
pectively), each based on their respective set of features. The bagging method is used to fit several
independent models and average their predictions in order to obtain a model with lower variance and
produce more reliable predictions based on classifiers from each feature set. The authors create nine
different subsets of tweets from the training set, including an equal number of samples for each class,
leading to nine instances of CL1 and CL2. The final prediction of the sample tests is calculated using
the average of the nine predictions concerning the classification algorithms like Logistic Regression and
Random Forest.

After the first-level classification, a second-level classification is made considering an agreement-
based retraining step (the fusion block in Figure 2.1), which is the key step of this framework. The step
combines the outputs of both classifiers CL1 and CL2 as follows: for each sample of the test set, the
outputs are compared, and depending on their agreement, the test set is divided into the agreed and
disagreed subsets. The agreed set elements are assigned to the agreed label (fake/real), assuming that
it is correct with high likelihood, and they are then used for retraining the best performing of the two first-level models to re-classify the disagreed elements. These retraining techniques create a model adjusted to the specific data characteristics and expect to predict the values of the samples more accurately than the first-level models CL1 and CL2, which did not initially agree.

The visualization component uses an annotated set of tweets called the Verification Corpus to show the end-users a set of feature distributions. From the user perspective, the tool asks for a URL or ID of a tweet of interest. The application presents the extracted tweet-based and user-based features and verification result (fake or real) for the tweet in the form of a color-coded frame green and red corresponding to true or false, respectively. The application also allows users to inspect each feature value showing the value distribution separately for fake and real tweets (side-by-side) and the percentage of this class’s tweets (fake or real) that have the same value for this feature. In this way, the user understands how the result is justified based on typical values that these features have on real and false tweets, as computed with respect to the verification corpus. Figure 2.2 presents the interface of the system.

### 2.3.2 XFake

Yang et al. (2019) proposed a system called Explainable Fake News Detector with Visualizations (XFake) to assist end-users in identifying news credibility. The authors designed three frameworks to detect the
news's credibility: MIMIC for attributes analysis, ATTN for semantic analysis, and PERT for linguistic analysis. The system provides explanations for the frameworks and examples similar to the news inserted to facilitate the visualization. The tool uses news data from a political fact-checking website named PolitiFact\(^9\) to train, tune, and evaluate. The **MIMIC Framework** is designed to analyze the news attributes. It employs a deep neural network to mimic the performance with tree ensemble models to simultaneously keep a good performance from neural networks and good explainability from tree ensemble models. The **ATTN Framework** is designed to analyze news statements from a semantic perspective using several techniques, including pre-trained word embedding, convolutional neural network (Krizhevsky et al., 2012), and self-attention mechanism (Vaswani et al., 2017). Self-attention is used to capture global relationships between different words efficiently, and it generates a weight matrix to help provide an instance-level explanation. Convolutional network helps to get explanations based on one-gram, two-gram, and three-gram analysis. The **PERT Framework** is designed for news statement analysis from a linguistic perspective. It employed six linguistic features such as adjective ratio, noun ratio, verb ratio, propn ratio (a subclass of nouns such as Google), sentiment score, normalized text length. For each item in the training set, PERT Framework extracts its linguistic features and train an XGBoost (Chen and Guestrin, 2016) classifier using these features to make predictions for a new item further. The prediction of the news' credibility is presented as a probability of being fake. The likelihood is calculated by combining MIMIC, ATTN, and PERT outputs in a weighted sum manner. The weights are tuned on a validation set, according to the performance of different frameworks. The explanation for the user is extracted from MIMIC (explained by key components), ATTN (explained by predictions from

\(^9\)https://www.politifact.com/
word/phrase attribution), and PERT (explained by linguistic features). Beyond this, XFake also explains predictions from a data perspective with supporting examples. These examples are generated from MIMIC and ATTN through the retrieval of training samples given corresponding explanations (i.e., key attributes or essential words/phrases). Figure 2.3 shows the architecture of XFake.

The visualization lies on three aspects. First, XFake presents histograms representing numerical values like prediction scores and attribute significance; second, it presents the outputs by highlighting important words/phrases with heatmaps to improve the explanation for word/phrases attribution; and third, it presents the ensemble trees which are capable of showing the overall structure and specific activated paths. Figure 2.4 shows the XFake’s interface for prediction and explanation.

2.3.3 NewsGuard

NewsGuard is a browser extension\(^\text{10}\) by NewsGuard Technologies to help to detect disinformation by reviewing and rating news and information websites based on nine journalistic criteria which assess

\(^\text{10}\)https://www.newsguardtech.com/
basic practices of **credibility** and **transparency**. The criteria are:

1. Does not repeatedly publish false content (credibility);
2. Gathers and presents information responsibly (credibility);
3. Regularly corrects or clarifies errors (credibility);
4. Handles the difference between news and opinion responsibly (credibility);
5. Avoids deceptive headlines (credibility);
6. Website discloses ownership and financing (transparency);
7. Clearly labels advertising (transparency);
8. Reveals who’s in charge, including possible conflicts of interest (transparency);
9. The site provides the names of content creators, along with either contact or biographical information (transparency);

The rating process starts with a NewsGuard analyst assessing the site’s content against the nine criteria and attributing a nutrition label for the site according to its reporting. Afterward, the website proprietor is asked to comment on the reporter’s rating to have the proprietor’s perspective. Next, the
site receives a red or green rating based on the nine journalistic criteria. NewsGuard also has other categories of rating, such as Satire for humor or satire sites, indicating that it is not a real news website and Platform for sites that host user-generated content that it does not vet. Figure 2.5 shows the interface of the extension of the news website cnn.com.

2.3.4 FactMata

FactMata11 is a free website that provides a quality score for online content to address disinformation and helps consumers think critically about what they are reading. FactMata uses natural language processing, and they also have a team of experts, including journalistic, researchers, and advocacy groups help to accurately train their AI.

The articles’ rating contains many signals such as clickbait, racism, deception, sexism, stance, hate speech, claim validation, toxicity, subjectivity, obscenity, arguments, insult, bias, threats, bot-generated content, rumors, and fake reviews. FactMata computes a score for the signals, explain and present to the users the most likely ones related to the news article. The tool also shows a progress bar classifying the article as harmful (red), questionable (yellow), and good (green) content. Figure 2.6 presents the tool’s interface with the feedback resulting from an article assessment.

11https://factmata.com/
2.4 Summary

Firstly, Section 2.1 focused on the linguistic aspects of online articles proposed by Caled and Silva (2020). These metrics helped to develop the linguistic indicators and analyze the linguistic aspects of a text in Portuguese. The evaluation of the metrics highlights the lack of linguistic resources for Portuguese, namely resources like subjectivity lexica that were translated from English to Portuguese using automatic tools, which may compromise the performance of the analyzed metrics. Despite the difficulties, it was possible to achieve initial hints for the disinformation detection tasks. Section 2.2 presented a work by Fuhr et al. (2018) that describes a range of indicators based on linguistic analysis, social network analysis, and content attributes and how they can be developed for English. Some of these indicators also helped in implementing context-based indicators for FactMe. Section 2.3 presented related software to understand how the indicators are computed and presented to users in similar software. This survey helped implement the tool’s metrics and indicators and gave an idea of how an interface of these systems can be.
3

FactMe

Contents

3.1 FactMe user interface ................................................................. 26
3.2 Computation of metrics and indicators ........................................... 28
3.3 Calibration of linguistic indicators ............................................... 31
3.4 Computation of the linguistic indicators and disinformation scores in FactMe ... 38
3.5 Implementation ................................................................. 40
3.6 Summary ............................................................................ 40

23
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 3.1 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, which are presented in Section 3.2.1.

- 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.

This chapter presents the implementation of FactMe. Section 3.1 shows the user interface of FactMe and how the indicators will be presented to the users. Section 3.2 describes the computation of linguistic metrics and context indicators of FactMe. Section 3.3 presents two datasets of news articles for calibration of linguistic indicators. Section 3.4 shows how the linguistic indicators score and disinformation score are computed in FactMe. Section 3.5 presents implementation details about the FactMe prototype and finally, Section 3.6 summarizes the contents of this chapter.

3.1 FactMe user interface

This section shows how users interact with FactMe. Figure 3.2(a) shows the homepage of FactMe. The homepage offers two alternatives for the article input. Users can provide a URL of a news website. This method uses Newspaper3k, 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. 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 (Figure 3.2(b)). 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 $\leq 33\%$), yellow ($33\% < \text{score} \leq 66\%$), 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

1https://newspaper.readthedocs.io/en/latest/
(a) User input page;

(b) User interface presenting indicators of disinformation for a news article;

Figure 3.2: FactMe interface.
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. Figure 3.2(b) shows the interface of a search result in FactMe.

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\(^2\) (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.

The interface of the navigational components can be found in Appendix D.

### 3.2 Computation of metrics and indicators

The indicators that are computed in FactMe are: **emotion**, **subjectivity**, **affectivity**, **polarity**, **behaviour and psychological**, **grammatical** as linguistic indicators and **source credibility** and **related tweets** as context indicators. The following sections will describe the computation of the linguistic metrics for each category of indicators (Section 3.2.1) and the context indicators (Section 3.2.2).

#### 3.2.1 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. Next, I will describe the datasets used for the computation of the metrics of each category of indicators that were presented in Section 2.1.

• **Affectivity metrics**- For estimating this category’s metrics, we use the Portuguese version of Affective Norms for English Words (ANEW) (Soares et al., 2012), a dataset with lexica rated by a human subject. The rating is made on a scale of 1 to 9, where a rating of 1 is denoted highly negative, and 9 denoted highly positive. Since this dataset’s size is small, Caled and Silva (2020)
enriched this lexicon with 14,000 English lemmas provided by Warriner et al. (2013), which were translated to Portuguese with the Google Translation API.

- **BP metrics** - To identify the terms of each BP process, we obtained the annotations provided by the Brazilian Portuguese Linguistic Inquiry and Word Count (LIWC) Dictionary (Filho et al., 2013), which is a dictionary that groups words into categories and can be used for sentiment analysis. LIWC was created by Pennebaker et al. (2001) to be used in a LIWC2015\(^3\), a language psychology analysis framework that has established strong links between linguistic patterns, and personality and psychological state.

- **Emotion metrics** - To identify each Ekman’s basic emotions, I used a Portuguese lexicon that was initially annotated by Strapparava and Mihalcea (2008) as a list of English words related to the six basic emotions and later was translated to Portuguese by Martinazzo (2010).

- **Polarity metrics** - Previous studies have organized sentiment terms into sentiment lexica, with the respective annotations on the polarity (positive, negative, or neutral) of those terms. This lead to the creation of the datasets SentiLex-PT (Silva et al., 2012) and OpLexicon (Souza et al., 2011), which have a set of adjectives, verbs, nouns, and a score assigned to each word. The positive and negative words can be identified using SentiLex-PT as the basic lexicon, complemented by the terms of OpLexicon. The scores are assigned based on a calculation on word distances, with polarity known as a priory. Same words may have different polarities depending on the domain in which it occurs. For example, in the case of “hot”, “The soup is still hot” can be considered a positive predictor, and “The champagne is hot” can be considered a negative one. This information can also be correctly processed with these datasets (Carvalho and Silva, 2015).

- **Subjectivity metrics** - Wilson et al. (2005) proposed the MPQA Subjectivity Lexicon for English words, which consists of a list of subjectivity clues containing words marked as strongly subjective or as weakly subjective. An immediate way to use the MPQA in Portuguese is to translate the terms with Google Translate API and extract the stem of the terms to estimate the subjectivity.

- **Grammatical metrics** - To perform PoS and Named-entity recognition (NER) in Portuguese we have CitiusTagger/CitiusNec (Garcia and Gamallo, 2015). This tool has the advantage of identifying entity names composed of more than one word, and it features more specific PoS tags such as person, gender, and number inflections values for determiners pronouns, verb tenses values, and specific punctuation types. This category can also group the degree of informality (or typographical error ratio) of a text. For this, we have the Python package Hunspell\(^4\) loaded with a Portuguese dictionary to identify the typographical errors.

---

\(^3\)http://liwc.wpengine.com/

\(^4\)https://pypi.org/project/hunspell/
Prof. Danielle Caled provided a Python library that receives as input a list of words and computes the metrics presented in Section 2.1 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. Section 3.4 will explain how the metrics will help in the disinformation detection task.

3.2.2 Computation of context indicators

- **Source credibility** - This indicator uses the ERC (Entidade Reguladora para a Comunicação Social), 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 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** - As stated in Section 2.2, 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, number of followers, number of likes, date and if it is a verified profile by Twitter. To search the tweets, I used Tweepy, a Python library that accesses Twitter API. I used the search method from Tweepy 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.

---

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

30
3.3 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\(^7\) and Fake.br Corpus (Monteiro et al., 2018), to compute the disinformation metrics. The datasets are described in Section 3.3.1. 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\(^8\) (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.

3.3.1 Datasets

I used two datasets, Polígrafo and Fake.br Corpus, with news articles described below:

- **Polígrafo** is a Portuguese online journalistic project that performs fact-checking to Portuguese news articles. It has a set of journalists responsible for classifying the news article, not evaluating the newspaper or journalist. Their method consists of (i) verifying the source, (ii) checking on the documentary sources that can solidify the checking process, (iii) listening to the authors giving them the right to explain, (iv) contextualizing information, and then (v) classifying the information on the following rating scale:
  - True: When the information is true.
  - True, but: When the information is true, but it needs to have a contextualization to be understood.
  - Inaccurate: When the contained elements that distort from the reality.
  - False: When the information is false.
  - Pepper on the tongue: It is only attributed if the information is scandalously false. It is the maximum degree of falsehood.

Polígrafo classifies a range of articles from social networks, newspapers, online articles, photos, and videos, and always justifies their decision and, whenever possible, shares links, videos, photos, documents, and other material to clarify their assessment. I gathered 50 text articles from

\(^7\)https://poligrafo.sapo.pt/
\(^8\)https://matplotlib.org/
different categories (sports, politics, entertainment) and various sources such as Facebook, different news organizations and labeled them as True or False. The collected true articles were previously classified by Polígrafo as “True” or “True, but”, and the collected false articles were classified by Polígrafo as “False” or “Pepper on the tongue”.

- **Fake.Br Corpus**: As presented in the related work (Chapter 2), this corpus introduced by Monteiro et al. (2018) 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.

### 3.3.2 Metrics distribution

This section assesses the distribution of the metrics of each category of indicators for true articles (represented in green color) and false articles (represented in red color). The distributions are compared considering Polígrafo and Fake.br Corpus datasets, disregarding outliers (some odd results that may be distant from the rest of the data). The goal of studying the distributions is to be able to compare true articles and false articles with the help of the metrics.

**Figure 3.3:** Boxplot of emotion metrics for Polígrafo and Fake.br Corpus. The X-axis represents the metrics for true (green) and false (red) labels and the Y-axis represents the emotion appeal of the text in form of a percentage.

Figure 3.3 presents the boxplot for each dataset which contains the **emotion metrics** (anger, disgust, fear, happiness, sadness, and surprise) representing the emotional appeal of the text. Figure 3.3(a) shows that the distribution of the metrics in Fake.br Corpus is similar in both labels. Particularly, the surprise lexica were not found for the true articles, but it were found for the false articles. All other metrics vary mostly the same in both labels however, the range of the true articles is greater than the
false articles, which means that the true articles have articles with more emotion lexica found. Figure 3.3(b) refers to the Polígrafo dataset where disgust and sadness metrics have a greater range for the true articles and, happiness and fear metrics a greater range for the false articles showing to have more emotion lexica found for that label. Anger metric showed to have three-quarters of the false articles with anger lexica found unlike true articles, which did not found any. In general, some metrics in Polígrafo, namely disgust, and anger, differentiated better the articles’ veracity.

Figure 3.4 shows the subjectivity metrics for each dataset, having the strong subjective metric on the left side and the weak subjectivity metric on the right side for each label. In both datasets, the strong subjectivity has a broader distribution in the true label than on the false label. However, the weak subjectivity has broader distribution in the true label in Fake.br Corpus, while in the Polígrafo dataset, the weak subjectivity lexica is more frequent in false articles. The Polígrafo shows to have articles with higher subjectivity rate than in Fake.br Corpus, thus the results do not help much to differentiate the article’s veracity in subjectivity metrics in any dataset.

I analyzed for each dataset, the five metrics of the affectivity indicator (average, standard deviation, maximum, minimum, and difference between maximum and minimum) of valence, arousal, and dominance dimension but, I only present the dominance dimension since the other dimensions have similar distribution for both datasets. Figure 3.5 shows that in Fake.br Corpus, the distribution of the majority of the label’s scores is different, e.g., standard deviation with the true articles have almost three-quarters above the highest score of false articles showing to have a higher affectivity rating. The same does not happen with the Polígrafo dataset, which has similar scores for both labels, not helping differentiate between the true and false articles.

Figure 3.6 presents the four polarity metrics distribution for both datasets. In Figure 3.6(a), we can
Figure 3.5: Boxplot of affectivity metrics. Presents the metrics of dominance. The X-axis represents the metrics for true (green) and false (red) labels and the Y-axis represents the rating of affectivity metrics in a scale of 1-9.

notice that the true label scores in Fake.br Corpus have a greater range, having one quarter (25%) of the true articles more positive and negative words than false articles. Figure 3.6(b) shows that Polígrafo has found more positive words for the false label and more negative words for the true label. Regarding Figures 3.6(c) & 3.6(d), Fake.br Corpus have much more positive and negative contrasts found with the false articles than the true articles, differentiating well the article’s veracity. Polígrafo only detected articles with 1 or 2 contrasts in both labels; this may be due to the dataset’s size being smaller than Fake.br Corpus. In general, positive and negative words metric did not help much, and the polarity contrast metric differentiated better the article’s veracity in Fake.br Corpus than Polígrafo.

Figure 3.7 presents the distribution of the BP metrics. Figure 3.7(a) shows that the medians are similar in both labels, but true labels are a little broader than the false ones showing to have higher and lower values. In Figure 3.7(b), we can see that the two-quarters of the values of perceptuality in false articles are above the values of the true label, cognitively shows that almost one-quarter of the values of the true label are higher than the false label and similarly, the biological metric has almost one-quarter of the false label values above the true label. All other metrics have a similar distribution, varying only the position of the quartiles. The distributions, in general, do not show a big variation between the labels for both datasets.

Figure 3.8 presents the distribution of the grammatical metrics. Figure 3.8(a) shows that some metrics differentiate the labels like the diversity metric where the median of the diversity words found in true articles is 80% and in the false articles 60%. Also, redundancy has a greater range on false articles. Regarding the content words (verbs, adjectives, and nouns), the true articles’ distribution is similar, though the overall range of the true articles is greater than the false articles’ distribution, having
higher and lower values. The informality and non immediacy lexica were not found in this corpus. In Figure 3.8(b), the majority of false articles have higher verbs ratio and lower adjectives ratio than the true articles. Also, some informality lexica were found in both labels, with true articles having a greater range. Overall, diversity and redundancy metrics in Fake.br Corpus differentiated better the articles’ veracity and verbs and adjectives ratio in Polígrafo, all other metrics did not help differentiate the false articles from the true articles.
3.3.3 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 Logistic Regression implementation used in this work was provided by scikit-learn library\(^9\). For the disinformation prediction task, I employed the following parameters: 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). Precision, in this context, measures how many false/true articles are, in fact, false/true and is defined as the fraction of relevant instances among the retrieved instances:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (3.1)
\]

Recall measures how accurately our model is able to identify false and true articles and is defined as the fraction of total relevant instances that were actually retrieved:

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (3.2)
\]

F1-Score is the harmonic mean of Precision and Recall to find an intermediate value between these two metrics and is defined as follows:

\[
F_1 - \text{Score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (3.3)
\]

Accuracy measures the overall correctness of the system and is defined by the fraction of correct predictions over total number of predictions:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (3.4)
\]

In the formulas, TP, FP, TN, and FN stand for true positives, false positives, true negatives, and false negatives, respectively.

Table 3.1(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 3.1(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.
Table 3.1: Indicators performance in forecasting news veracity using Logistic Regression.

3.3.4 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.

3.4 Computation of the linguistic indicators 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 to
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 Section 3.3.3). 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. Figure 3.9 shows the computation of the indicator's score and
the combined indicator.

3.4.1 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 Section 3.1. 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.

3.5 Implementation

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. Since the application is built with Flask, I used
Gunicorn (Green Unicorn), which is a WSGI implementation that is commonly used to run Python web
applications. A Python file was created to serve as a WSGI entry point to the application and a Unix
socket file to receive the requests. To receive web requests to the socket file, I used the Nginx web
server that can be used as a reverse proxy. With this configuration, the web application is hosted in the
following link: http://194.210.120.9/.

3.6 Summary

This chapter presented how FactMe was implemented. I started by describing the information process-
ing of FactMe and showing how disinformation will be presented to the end-user. I then described how
the linguistic metrics and the context indicators are computed. I used two datasets with news articles,
Polígrafo and Fake.br Corpus, and created a dictionary with the computed metrics of the articles for each dataset. Using the dictionary created, I made a calibration process using boxplots and Logistic Regression algorithm to evaluate each linguistic indicator category and the datasets. Furthermore, I also explained how disinformation detection could be formulated as a classification task presenting the input and training set to compute the linguistic indicator's score and the disinformation score in FactMe.
43
This chapter describes the methodology used for the evaluation of FactMe as well as the obtained results, analysis, and conclusions. Chapter 4.1 presents the disinformation indicators performance evaluation and Chapter 4.2 presents the FactMe user interface evaluation.

4.1 Evaluation of the disinformation indicators

To assess the disinformation indicators’ performance, I created a page in the navigational bar of FactMe called “evaluation”, where the users could assess the FactMe prototype and record their assessments. Appendix A has the screenshot of the evaluation page along with the questionnaires.

4.1.1 Characterization of users

The FactMe web application uses cookies, but it does not keep track of any data or IP addresses from the user, aside from the answers they provide. The evaluation has a single requirement: the respondent must be fluent in Portuguese to read the articles. Besides that, it does not need any background or skill.

4.1.2 Questionnaire

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, used in Section 3.3.1. 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 2 versions of questionnaires to be answered alternatively in order to compare them later. Figure 4.1 shows the evaluation topology. Both versions will be described as follows:
Figure 4.1: Alternative questionnaires for the evaluation of the FactMe disinformation indicators.

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, the user is 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. Table 4.1 presents the questionnaire made for each variant.

There are in total six questionnaires, each one having a different news article. Each version has these six questionnaires with the same articles in each questionnaire. Version 1 will show each questionnaire without indicators (variant A) and then with indicators (variant B) and, version 2 will show the same questionnaire with indicators (variant B). The user answers the questionnaire one at a time of 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
Table 4.1: Questionnaire of variant A and variant B to evaluate the disinformation indicators. The questions related to each variant are marked with a cross.

<table>
<thead>
<tr>
<th>Questions</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Does the title represent the content of the article?</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2  Why do you think the title does not represent the content of the article?</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3  Is the title clickbait?</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4  If you think the headline is clickbaity, what made you think it is?</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>5  Is the article subjective (it is expressed an opinion)?</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>6  How do you classify this news?</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7  Did the indicators made an impact on the credibility of the article?</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>8  Which indicator made the biggest impact?</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

( 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.

### 4.1.3 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. The green and red colors correspond to the true and false news articles, respectively, the Y-axis corresponds to the number of answers, and the X-axis displays the options of the question.

Figure 4.2 presents the distribution of the answers to question one, two, and three, which are related to the title of the article. Figure 4.2(a) shows the results for the **first question** (does the title represent the content of the article?). 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.

Figure 4.2(b) presents the results of **second question** (why the title does not represent the content of the article?) where the X-axis in the chart corresponds to the following answers:

- It is representative (A)
(a) Does the title represent the content of the article?

(b) Why the title does not represent the content of the article?

(c) Is the title clickbaity?

Figure 4.2: Distribution of the answers related to the title of the article.

- Title is on a different topic than the body (B)
- Title carries only a little information about the body (C)
- Title takes a different point of view than the body (D)
- Title overstates/understates claims or conclusions (E)

As we saw in Figure 4.2(a), a large majority answered somewhat or completely representative. The correlation of this question with the first question is reflected in the answers, with most answers being A (it is representative) in both versions. Options C and E were also selected from a few respondents. The true articles had mostly a representative title but as well as the false one.

Figure 4.2(c) presents the third question (Is the title clickbaity?) where 1 corresponds to not clickbaity, and 5 corresponds to very clickbaity in X-axis. Variant A shows that the true articles are not that
clickbait 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. Table 4.2 presents the answers for true and false articles and a translation to English for the Portuguese answers.

We can observe 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.

Figure 4.3 presents the distributions of answers to questions five and six related to the article’s subjectivity and classification. Figure 4.3(a) presents the results of the fifth question (Is the article subjective?) where 1 corresponds to not subjective, and 5 corresponds to very subjective in X-axis. It can be seen 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 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

<table>
<thead>
<tr>
<th>Original answer</th>
<th>English translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uso de nomes mais conhecidos e título muito subjectivo em relação à notícia</td>
<td>Use of known names and very subjective title in relation to the news</td>
</tr>
<tr>
<td>Because the title claims something that isn’t confirmed on the body</td>
<td>-</td>
</tr>
<tr>
<td>Article is not specific</td>
<td>-</td>
</tr>
<tr>
<td>Uso de nomes mais conhecidos</td>
<td>Use of known names</td>
</tr>
</tbody>
</table>

Table 4.2: Answers to the question "What makes you think the title is clickbaity?".

true articles

<table>
<thead>
<tr>
<th>transmite escândalo</th>
<th>transmit scandal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Referir que é uma proposta milionária faz com que as pessoas queiram saber qual o valor</td>
<td>To mention that it is a millionaire proposal makes that people want to know what is the value</td>
</tr>
<tr>
<td>Conclui o título quando não se concluir ainda</td>
<td>It shocks the audience that reads the title</td>
</tr>
<tr>
<td>Questões político-religiosas são sempre motivo de discussão e debate. Provoca essa emoção nas pessoas.</td>
<td>Political-religious issues are always a matter of contention and debate. It causes this emotion in people</td>
</tr>
<tr>
<td>The articles’ point is completely different than the titles</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

Figure 4.3(b) shows the results for the sixth question (how do you classify this news?), one of the most relevant. 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.

Figure 4.4 presents the distributions of answers to questions seven and eight related to the impact of the disinformation indicators. For this pair of questions, I only analyzed assessments with indicators that correspond to variant B in both versions. Figure 4.4(a) shows the results for the seventh question (did the indicators made an impact in the credibility of the article?) which 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.
Did the indicators made an impact in the credibility of the article?

Which indicator made the biggest impact?

Figure 4.4(b) shows the result for the eighth question (which indicator made the biggest impact?). 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.
4.2 Evaluation of the user interface

4.2.1 Questionnaire

To evaluate the user interface, I used the QUIS (Questionnaire For User Interaction Satisfaction), a tool developed to assess users' subjective satisfaction with specific aspects of the interface in a standardized, safe and valid way. QUIS was originated from a team of researchers in the HCIL at the University of Maryland and is currently in version 7.0 (Chin et al., 1988). 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.

The questionnaire is attached in Appendix B. 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.

4.2.2 Results

The assessment was made by 11 people. Table 4.3 presents the questions related to the six aspects, their results, 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, 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.
<table>
<thead>
<tr>
<th>Questions</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>Average</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Software reaction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.73</td>
<td>9</td>
</tr>
<tr>
<td>Bad - Good</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>8.73</td>
<td>9</td>
</tr>
<tr>
<td>Difficult - Easy</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>8.82</td>
<td>9</td>
</tr>
<tr>
<td>Frustrating - Satisfying</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8.73</td>
<td>9</td>
</tr>
<tr>
<td>Rigid - Flexible</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8.36</td>
<td>8</td>
</tr>
<tr>
<td>Inadequate - Adequate</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9.00</td>
<td>9</td>
</tr>
<tr>
<td><strong>Screen</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>8.55</strong></td>
<td>9</td>
</tr>
<tr>
<td>Font and size</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8.82</td>
<td>9</td>
</tr>
<tr>
<td>Highlighting on the screen</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8.00</td>
<td>9</td>
</tr>
<tr>
<td>Organization of the information</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8.64</td>
<td>9</td>
</tr>
<tr>
<td>Sequence of screens</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8.55</td>
<td>9</td>
</tr>
<tr>
<td>Screen items easy to find</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8.73</td>
<td>9</td>
</tr>
<tr>
<td><strong>Terminology and system information</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>8.40</strong></td>
<td>8</td>
</tr>
<tr>
<td>Use of terms throughout system</td>
<td>8</td>
<td>9</td>
<td>6</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>8.46</td>
<td>9</td>
</tr>
<tr>
<td>Messages that appear on screen</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>8.37</td>
<td>8</td>
</tr>
<tr>
<td>Position of the messages</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>8.37</td>
<td>8</td>
</tr>
<tr>
<td>Computer keeps you informed about what you are doing</td>
<td>8</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8.18</td>
<td>8</td>
</tr>
<tr>
<td>Error messages</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8.64</td>
<td>9</td>
</tr>
<tr>
<td><strong>Learning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>8.43</strong></td>
<td>9</td>
</tr>
<tr>
<td>Learning to operate the system</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8.55</td>
<td>9</td>
</tr>
<tr>
<td>Exploring by trial and error</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>6</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>8.64</td>
<td>9</td>
</tr>
<tr>
<td>Remembering terms and use of commands</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8.64</td>
<td>9</td>
</tr>
<tr>
<td>Tasks can be performed in a straight-forward manner</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8.64</td>
<td>9</td>
</tr>
<tr>
<td>Help messages</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8.64</td>
<td>9</td>
</tr>
<tr>
<td>Supplemental reference materials</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8.36</td>
<td>9</td>
</tr>
<tr>
<td><strong>System capabilities</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>8.13</strong></td>
<td>8</td>
</tr>
<tr>
<td>System speed</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>8.27</td>
<td>8</td>
</tr>
<tr>
<td>System reliability</td>
<td>8</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8.18</td>
<td>8</td>
</tr>
<tr>
<td>Correcting your mistakes</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>6</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>8.09</td>
<td>8</td>
</tr>
<tr>
<td>Experienced and inexperienced users are considered</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>8.00</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td><strong>Usability and user interface</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>8.61</strong></td>
<td>9</td>
</tr>
<tr>
<td>Use of colors</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>8.45</td>
<td>9</td>
</tr>
<tr>
<td>System response to errors</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>6</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>8.36</td>
<td>9</td>
</tr>
<tr>
<td>System messages</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>8.82</td>
<td>9</td>
</tr>
<tr>
<td>Feedback of the system</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8.82</td>
<td>9</td>
</tr>
<tr>
<td><strong>Global average</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>8.48</strong></td>
<td>9</td>
</tr>
</tbody>
</table>

Table 4.3: Results of FactMe evaluation with QUIS.
4.3 Summary

This chapter covered two evaluations made on FactMe. An evaluation of the impact of disinformation indicators where the focus was to evaluate the tool’s effectiveness with two variants, one with indicators and the other without them presented in different approaches. Also, an interface evaluation to measure user satisfaction was made using QUIS.

The evaluation showed that the indicators were less effective to help users evaluate the article’s title but made a difference in the articles’ classification and a better prediction of the articles’ veracity when the indicators were presented in variant B. Also, notice that showing the indicators first in Version 2 was slightly better. Some indicators like subjectivity and emotion made a better impact than others, but the source credibility indicator made the most significant impact.

Regarding the interface, FactMe received very positive feedback, majorly in all QUIS factors, having an average above 8 and a median of 8 and 9 in almost all ratings. Also, the open questions showed the system to be simple and easy to use. Some improvements were also mentioned, such as a better Homepage and some more help messages to facilitate the tool’s usage. It is important to recognize that the population that evaluated the user interface was small and biased. It is noteworthy that the grammatical indicator was implemented after the evaluation phase, so it was not considered in both evaluations. Since it takes a little more computational time, the system speed factor in QUIS could have been different.
5

Conclusion

Contents

5.1 Contributions .......................................................... 59
5.2 Future work ............................................................. 59
The consumption of fresh news content has changed over the last years with the growth of social media and online news sources. This growth can expose users globally and enable them to share knowledge, which can also contribute to spreading misinformation. Fake news can travel significantly farther, faster, deeper, and more broadly than truthful news in all categories of information (Vosoughi et al., 2018). In such circumstances, it is vital to prepare news consumers to stop consuming misinformation by investing in, for example, automatic tools, fact-checking sites, and education. FactMe addresses the problem as an online visualization tool providing information quality indicators for Portuguese news articles.

This chapter concludes the dissertation, reviewing in Section 5.1 the work done and, in Section 5.2 some possible ideas to extend and improve FactMe.

5.1 Contributions

In this work, I have surveyed previous research on content-based (linguistic) and context-based metrics, including some misinformation 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 misinformation 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 misinformation consumption.

5.2 Future work

FactMe is a first version prototype that can be improved. Since changes in the news consumption world can occur fast, some indicators of FactMe can be compromised over time, particularly the related tweets and source credibility. An improvement can be made to FactMe by removing or adding more indicators, such as some indicators from Section 2.2. For example, the source credibility indicator, which showed to
be the most powerful in FactMe, is effective, although it depends on the agencies registered on the ERC sheet. Another alternative would be to use the Pagerank algorithm by evaluating the importance and relevance of a webpage based upon the number of quality and credibility of links to the page in question.

FactMe is implemented to be used without any authentication. It is also possible to improve the tool by adding an authentication system, with a review option from the consumer side to get feedback for the tool's judgment.

Lexica like ANEW for the affectivity metrics had to be translated with automatic tools. It may have compromised the performance in detecting disinformation. More powerful lexica for Portuguese could be needed in order to achieve better results.

The collection of articles evaluated from the Polígrafo fact-checking site was challenging since the news articles were not explicitly shared. I created a small dataset by finding the articles one by one on the web. Fake.br Corpus was more effective to use than Polígrafo, but the articles are not recent. Having a good accuracy in disinformation detection tasks with one corpus may not be enough. The disinformation indicators are expected to be validated with different corpora. Additionally, the corpus can also differentiate news articles by their categories, e.g., sports, health, politics, etc. With this differentiation, it can be possible to perform disinformation detection in different categories and achieve better results.

Current datasets have the articles classified with binary labels such as true or false. One more idea can be not to classify articles as true or false labels but to use a more fine-grained information classification. Classifications like satire, biased, political can be added to a corpus in order to differentiate closely related from truly fake articles.

Furthermore, the evaluation of FactMe, namely the user interface evaluation, did not identify the users and was highly biased. An evaluation can be made with a larger population to get more precise details about the interface of FactMe and identify the users to consider the results according to them.
Bibliography


Evaluation questionnaire of indicators

Evaluation
This assessment contains a way to be evaluated which can be an article from the Social Media or from a news website which the veracity is known. Your goal is to read the article and answer the questions associated with the credibility of the article.
Next you will find the same article but this time are presented the indicators of disinformation provided by the tool and you will have to answer the same questions taking into account the indicators.
The assessment aims to evaluate the indicators proposed by this tool and assess the usefulness and impact on the answers obtained.

Figure A.1: Evaluation without indicators questionnaire (1).
1. Does the title represent the content of the article?
   - Completely Unrepresentative
   - Somewhat Unrepresentative
   - Somewhat Representative
   - Completely Representative

2. If you selected unrepresentative, why do you think the title does not represent the content of the article?
   - It's representative (if you selected representative)
   - Title is on a different topic than the body
   - Title carries only little information about the body
   - Title takes a different point of view than the body
   - Title overstates/understates claims or conclusions.

3. Is the title clickbait (form of false advertisement to attract attention and to make users read the news)?
   Rank on a scale of 1 (Not clickbait) to 5 (Very clickbait)
   - 1
   - 2
   - 3
   - 4
   - 5

**Figure A.2:** Evaluation without indicators questionnaire (2).

4. If you think the headline is clickbait, what makes you think it is?
   Answer:

5. Is the article subjective (it is expressed an opinion)?
   Rank on a scale of 1 (Not subjective) to 5 (Very subjective)
   - 1
   - 2
   - 3
   - 4
   - 5

6. How do you classify this news?
   - True
   - Dubious
   - Fake

**Figure A.3:** Evaluation without indicators questionnaire (3).
Figure A.4: Evaluation with indicators questionnaire (1).

Figure A.5: Evaluation with indicators questionnaire (2).
Figure A.6: Evaluation with indicators questionnaire (3).
Evaluation questionnaire of user interface (QUIS)
Avaliação de interface de FactMe

O FactMe foi implementado como um projeto de tese para o Mestrado em Engenharia e Informática e de Computadores do Instituto Superior Técnico. FactMe é um protótipo que apresenta indicadores de qualidade da informação de forma explicável para notícias em português para ajudar a detetar desinformação e auxilia os consumidores a julgar a credibilidade dos artigos.

O objetivo desta avaliação é utilizar a ferramenta FactMe e avaliar a interface de utilizador usando o QUIS (Questionnaire for User Interface Satisfaction), um questionário utilizado para medir a satisfação de um utilizador quanto a usabilidade da ferramenta.

É pedido que submeta alguns artigos em português na Homepage ou que consulte um dos artigos já avaliados anteriormente no separador Fact-checks para avaliar a interface de utilizador e responder a este questionário.

Não vai ser guardada qualquer informação pessoal neste questionário.

Website de FactMe: http://194.210.120.9/HomePage

Muito obrigado.

Figure B.1: Questionnaire description.
Figure B.2: Reaction to software and screen aspects.
(a) Terminology & system information

(b) Learning

Figure B.3: Terminology & system information and learning aspects.
(a) System capabilities  
(b) Usability & user interface

Figure B.4: System capabilities and usability & user interface aspects.

Figure B.5: Open questions.
Prototypes
Figure C.1: Paper mockup for user input of FactMe.

Figure C.2: Paper mockup of FactMe presenting indicators for a news article.
Figure C.3: Digital mockup for user input of FactMe.

Figure C.4: Digital mockup of FactMe presenting indicators for a news article.
FactMe Navigational components interface
Figure D.1: Fact checks component interface.

Figure D.2: How it works component interface.
Figure D.3: About Me component interface.