Submission date: 25/11/2019
Resubmission date: 30/04/2020
Resubmission date: 28/06/2020
Camera ready submission date: 04/07/2020

1st round notification: 09/03/2020
2nd round notification: 09/06/2020
3rd round notification: 01/07/2020
Available online: 06/07/2020
Publication date: 31/07/2020

Section: extended version
Extended versions of papers presented at the XV Brazilian Symposium on Information Systems (SBSI 2019)

A Multi-label Classification System to Distinguish among Fake, Satirical, Objective and Legitimate News in Brazilian Portuguese

Janaína Ignácio de Morais[1], Hugo Queiroz Abonizio[1], Gabriel Marques Tavares[1], André Azevedo da Fonseca[2], Sylvio Barbon Jr.[1]

[1]Computer Science Department – State Univerisity of Londrina (UEL)
Londrina, Paraná – Brazil

[2]Comunication Department – State University of Londrina (UEL)
Londrina, Paraná – Brazil

{janainam, hugo.abonizio, gtavares, andre.azevedo, barbon}@uel.br

Abstract. *Currently, there has been a significant increase in the diffusion of fake news worldwide, especially the political class, where the possible misinformation that can be propagated, appearing at the elections debates around the world. However, news with a recreational purpose, such as satirical news, is often confused with objective fake news. In this work, we decided to address the differences between objectivity and legitimacy of news documents, where each article is treated as belonging to two conceptual classes: objective/satirical and legitimate/fake. Therefore, we propose a DSS (Decision Support System) based on a Text Mining (TM) pipeline with a set of novel textual features using multi-label methods for classifying news articles on these two domains. For this, a set of multi-label methods was evaluated with a combination of different base classifiers and then compared with a multi-class approach. Also, a set of real-life news data was collected from several Brazilian news portals for these experiments. Results obtained reported our DSS as adequate (0.80 f1-score) when addressing the scenario of misleading news, challenging the multi-label perspective, where the multi-class methods (0.01 f1-score) overcome by the proposed method. Moreover, it was analyzed how each stylometric features group used in the experiments influences the result aiming to discover if a particular group is more relevant than others. As a result, it was noted that the complexity group of features could be more relevant than others.*

Keywords. *Fake News, Decision Support System, Text Mining and Multi-Label*

1. Introduction

Nowadays, the way of consuming, interpret, and process citizens' news has changed substantially, mainly due to the ease of communication existing on social networks. Through the 20th century, it was common that information circulating in mass media, where the

Cite as: Morais, J. I., Abonizio, H. Q., Tavares, G. M., Fonseca, A. A. & Barbon, S. (2020). A Multi-label Classification System to Distinguish among Fake, Satirical, Objective and Legitimate News in Brazilian Portuguese. iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 13(4), 126-149. DOI: 10.5753/isys.2020.833

news attended classic journalism criteria - such as timeliness, social relevance, and integrity [Shu et al. 2017]. Naturally, this tradition was not free from ideological biases or even deliberate distortions from various factors - from pure sensationalism, which aimed to increase sales (or audience) through exaggerated emotional appeal; to the overpowering or defamatory news commissioned by politicians interested in interfering in public opinion; or even mere negligence in the newsgathering, the result of unfavorable working conditions in a context of industrial production of journalism.

Besides, by holding down the power to select and broadcast news, the media played a decisive role in the political culture in which they were inserted, as they were able not only to attribute meanings to facts but to determine which facts are relevant to the public debate and which should be ignored. In other words, the media gained the prerogative to define the themes that became journalistic coverage and, consequently, public debate. The classical agenda-setting theory explains this dynamic [McCombs and Shaw 1972]. The only possible reaction of counter-hegemonic voices was the alternative, more restricted-circulation media.

However, the credibility of the media was built precisely from the commitment of several media companies to the practices of professional journalism, resulting in readers who entrust the media the task of processing the enormous volume of information about reality and presenting it in an assimilable subset of daily news [Lazer et al. 2018]. This dynamic is understood from the theory of gatekeeping. ”Gatekeeping is the process of culling and crafting countless bits of information into the limited number of messages that reach people each day, and it is the centre of the media’s role in modern public life” [Shoemaker and Reese 2013].

Nevertheless, in the context of access to information through social networks, the mediation of news, especially in the final stage of consumption, is no longer carried out by journalists and professional editors. Despite the discourse that social networks abolished the gatekeeper - since each member of the virtual community, in exercising their alleged freedom of choice, would have been promoted to the status of an editor of their content - in practice many argue that this dynamic only changed its place. Therefore, instead of professional journalists, algorithms have become primarily responsible for selecting and distributing information that reaches individual consumers [Pariser 2011].

This action has been coupled with the proliferation of amateur sites that obtain high profitability with online traffic, through easy access to digital ad programs such as Google AdSense. Driven by social networks and enhanced by data analyses that indicate the tastes, prejudices and predispositions, a multitude of websites is dedicated to producing content for easy and quick spread, without any compromises with issues unrelated to their profitability. Thus, users tend to receive and consume a massive volume of information of dubious origin.

The advent of social networks also caused a crisis in the ability to hierarchize information. Mark Zuckerberg, the creator of Facebook himself, unwittingly stated the terms of the problem in a famous statement: ”A squirrel dying in front of your house may be more relevant to your interests right now than people dying in Africa” [Pariser 2011]. That is, more than to equate issues with such unbalanced dimensions, distributing famil-

iar frivolities and international catastrophes to the same extent, the algorithm that organizes the Facebook timeline degrades the ethical operations of classifying information, imposes obstacles to the sizing of facts, and implodes the boundaries between relevance and irrelevance, dumbing down users in the field of media criticism by inducing them to a narcissistic, distracted and distorted interpretation of the content. A survey released in September 2018 by Ipsos[1], about 62% of the Brazilian population, believes in the fake news, which shows how aggravating this problem inserted in our society.

Studies in the field of media literacy [Kress 2003] have been seeking for years to educate the public on particularities of messages in media. The growing complexity of the media ecosystem requires the formation of readers and spectators capable of understanding the diversity of factors that condition the production of information. Nevertheless, to establish a critique of the contents, it is essential to formulate a reflection on the media's language. Without such instruction, users have fewer resources to discern between the misleading language of a fake news website and the ambiguous language of a humorous site, for example.

Sensacionalista[2] is one of the most popular humour sites in Brazil. Amidst an explicit parody of journalistic language, editors create comic stories, including real personalities, and inspire social criticism through irony. For this, Sensacionalista - who even mocks the name - is not properly defined as a fake news site, since the stated objective is humour.

Though, inattentive readers - and without training in media literacy - frequently are confused when interpreting funny texts as real stories. As irony is a complex language feature, the joke is not always obvious. Moreover, the writers' ability to construct the parody in the form of news seeks to extract humour through allegory precisely. The differences are usually quite subtle.

Already fake news is better defined as "news articles that are intentionally and verifiably false and could mislead readers" [McCombs and Shaw 1972]. The most intense international debate on the subject occurred mainly after the results of the United States election. Furthermore, according to a survey released in October 2018 by Datafolha[3], most of the news propagated in the last Brazilian elections come from social networks like Facebook.

Furthermore, there are very particular characteristics with the irony present in the texts, where we can visualize negative or opposite feelings in some affirmations. Regarding the politics, besides irony, we can correlate the sentiment analysis with the defeat of the people fronting a particular party, where it influences the results of the research [Tayal et al. 2014].

Although there are various Text Mining (TM) works producing state of the art performance addressing the issue of fake news detection based on its textual content, we

[1]https://www.ipsos.com/pt-br/global-advisor-fake-news

[2]https://www.sensacionalista.com.br/

[3]http://datafolha.folha.uol.com.br/opiniaopublica/2018/10/1983765-24-dos-eleitores-usam-whatsapp-para-compartilhar-conteudo-eleitoral.shtml

believe that a piece of news can carry multiple conceptual classes. At the state of the art, there are several works of Text Mining (TM) with good results based on its textual content. However, we believe that a piece of news can carry multiple conceptual classes. Therefore, an additional challenge is posed to traditional TM towards evolving the task into multi-label textual classification.

Standard single-label classification addresses the induction of a model from a set of examples associated with a unique label l from a set of disjoint labels L, $|L| > 1$. If $|L| = 2$, we have a binary classification problem. Alternatively, if $|L| > 2$ it is a multi-class classification scenario. Already in the multi-label classification, examples are related to a set of labels $Y \subseteq L$. As declared, we consider that a news article is associated with a set of possible conceptual class $Y \subseteq L$.

In the context of misleading news as a multi-class problem, we have $|Y| = 2$ with conceptual classes of $y_1 =$ "fake/legitimate" and $y_2 =$ "satirical/objective". The labels are $L = \{$objective-legitimate, objective-fake, satirical-fake, satirical-legitimate$\}$

Hence, this study aims to propose and validate a pipeline for text mining with a multi-label classification of news embedded in a Decision Support System (DSS) of news legitimacy. The conceptual classes are described to its falsity (fake/legitimate) or its objectivity (objective /satirical).

The secondary contributions of this work are:

1. Present our real-life multi-labeled news dataset;
2. Identify the best machine learning algorithm and textual features in our multi-label news scenario;
3. Propose new textual features and evaluate the impact of them in the results classification;

This paper is an extended version of a previous work [de Morais et al. 2019], including the study of significant features groups and if it is possible to obtain the same result compared with the initial experiments using fewer features. The remainder of this paper is organized as follows. Section 2 presents an overview of the related research. Section 3 presents a proposed approach, showing the pipeline and feature extraction used in this research. Section 4 presents a description of the methods and model evaluation. Section 5 presents the results and discussion. In the last, we discuss the results achieved by our experiments. The conclusion and future work are presented in section 6.

2. Related Work

In recent years, the detection of fake news has been the subject of several works on state-of-the-art literature. Most of them can be split into two main categories: news content-based and social context-based [Shu et al. 2017]. In this work, the focus is on the category based on news content, where this content can be analyzed in order to decide its falsity and objectivity.

Recent work by [Shu et al. 2017] focuses on a comprehensive analysis of fake news detection in social media, taking into account characteristics such as fake news concepts in traditional and social media. A binary classification was also used, reaching

a list of significant attributes such as title, text body, possible images, among others. For this, the authors suggested possible ways of solving through machine learning techniques, leaving open ways of exploring this problem with data mining.

To detect fake news based on your content, [Allcott and Gentzkow 2017] proposed the usage of Natural Language Processing (NLP). The authors consider only the textual content of the news and processed this contend with the Recurrent neural network (RNN) and the Gated Recurrent Unit (GRU). The result did not completely exploit the dataset, which resulted in a low comprehensive model.

The authors [Singhania et al. 2017], aiming to improve fake news prediction in news stories, compared a 3-level hierarchical classifier (words, phrases, and headlines). The author's classification proposal concerning different aspects of the model achieved proper results. However, its contribution is limited and consider only binary classification.

The most common source of fake news is online social networks. In this scenario, [Shao et al. 2017] identified potential bots origin of spreading fake news on Twitter. The authors proposed a tool that recognizes the dissemination of misleading information by tracking those accounts responsible for the initial spreading of news and some related patterns. However, a critical discussion of this work is when a news article reaches ordinary people who believe in the content and share it with friends and followers, creating a cycle. This phenomenon, on a large scale, compromises the identification of real conceptual class.

A study about user behavior using Twitter was proposed by [Ruchansky et al. 2017]. The authors present a model that combines three characteristics (article text, user response, and unique user) to predict fake conversations. The results contributed to represent users and articles towards identifying significant sources of risk.

Nevertheless, only pure detection of fake news is a challenging task since fake news is not yet entirely understood, as seen by [Ruchansky et al. 2017]. According to [Rubin et al. 2016], sarcastic and ironic news can also be a form of fake news. With this, this type of news can be confused with fake news, and depending on how it came to the consumers; there is a big possibility to interpret them as legitimate news. The twitter analysis was the theme of [Tayal et al. 2014] work, where tweets were analyzed based on two proposed measurements, the first to identify a given tweet as sarcastic and the second to detect the polarity in sarcastic political tweets.

Similarly, work by [González-Ibánez et al. 2011] created a search engine for sarcastic tweet content. Intending to examine the impact of lexical and pragmatic factors, the authors made a comparison between Machine Learning techniques (Support Vector Machine and Logistic Regression) and human beings in sentiment classification. The overall accuracy was low due to difficulties of sarcasm classification in both cases. Despite that, the human being won with a little difference in this case.

The authors [Poria et al. 2016] proposed the use of Convolutional Neural Network (CNN) to extract feelings, emotions, and personality in the detection of sarcasm, making a subjectivity analysis from Twitter. The results obtained exceeded state of the art; though,

it is worth considering the experimentation was conducted with a single news source.

More recent studies worked on increasing of the interpretability of fake news detection systems. FakeNewsTracker [Shu et al. 2019b] and dEFEND [Shu et al. 2019a] are examples of methods to make those decision support systems explainable and enable the visualization of decision results.

Taking into account proposals related to TM, some researches in the literature dealt with multi-label classification [Ishita et al. 2010, Bhowmick 2009, Li et al. 2016, Li et al. 2016, Almeida et al. 2018], and a significant part of them devoted to sentiment analysis and multiple topic classifications. It is important to mention the contribution of [Almeida et al. 2018] in comparison to a wide range of techniques giving relevant insights about the bias of multi-label techniques.

Most of the related work presented was based on binary classification where news is only true or false, supported by textual and non-textual features, and focused on specific sources, e.g., Twitter. However, our DSS is based only on textual features extracted from the news by a straightforward text mining pipeline. We evaluated our proposal with different news sources to reduce the bias of a single portal. Additionally, our proposal discusses the multi conceptual class of single news, as stated in the presented multi-label definition.

3. Proposed Approach

The Decision Support System proposed in this work uses a pipeline for classifying news documents; for this, it is used stylometric features extracted from the text. This approach aims to classify documents into two conceptual classes: fake/legitimate and satirical/objective, which makes a total of 4 possible class combinations (objective-legitimate, objective-fake, satirical-legitimate, and satirical-fake).

The whole DSS could split into two parts: the creation of the DSS and your execution to obtain a prediction. Figure 1 shows the DSS creation steps, where the model built with data raised previously. Figure 2 refers to the process of executing the created system either on a validation set to evaluate the result or on a production environment with new data. Figure 1 and Figure 2 represents the phases of this pipeline.

Beyond of creation and execution of this DSS, we also focus on analysis from how stylometric features extracted from the text can influence the result of document classification. So, this analysis can split into two parts: the obtained and split of group features and analysis of these features, where each group of features is separately analyzed and combined with other groups to find out if a particular group of features tends to obtain good results without other groups.

3.1. Text processing

Regarding DSS creation, the first step (1) is pre-processing of the raw dataset, where are performed the text cleaning [Almeida et al. 2018], Part-Of-Speech (POS) tagging [Collins 2002] and stopword removal [Igawa et al. 2014]. During this step, useless spaces and special characters are converted and then tokenized. Each token is assigned with a POS label, and stopwords are removed. Since stopwords have a high frequency across

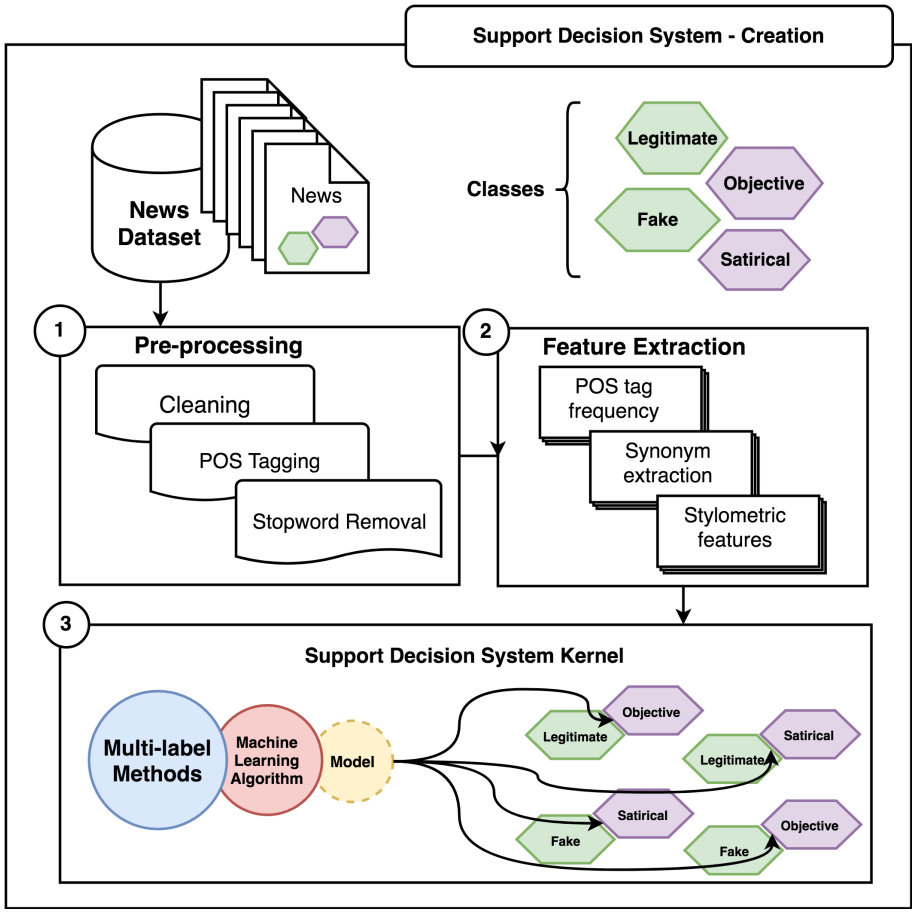

Figure 1. Creation of Decision Support System for detection of fake, legitimate, satirical or objective news [de Morais et al. 2019].

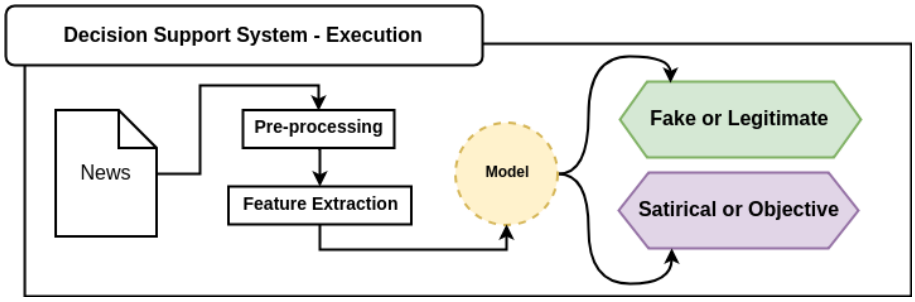

Figure 2. Execution of Decision Support System for detection of fake, legitimate, satirical or objective news [de Morais et al. 2019].

documents, they are considered noise on text data with little discriminative power with this, and its removal usually improves the performance of the model [Saif et al. 2014], those words are not useful for our purposes.

In the second step (2), Feature Extraction, the frequency of POS tags, the average number of synonyms per term, and other stylometric features are obtained (more details in Section 3.2).

After the processes occurred on step (2), feature vectors are generated, and each instance is equivalent to a document from the raw dataset. Each instance has two labels, one for each conceptual class, and is used to induce the decision model. Then, on the third step (3), a machine learning algorithm is used to create a prediction model using a given multi-label method with those feature vectors. This model is the kernel of the decision process, where the machine learning model extracts patterns to distinguish the classes fashioned by a multi-label domain.

As we are classifying documents that can be legitimate or fake and satirical or objective, the multi-label approach is more suitable than a simple multi-class classification. Instead of belonging to a single category, a document is labeled as either one class and another at the same time, e.g., fake *and* satirical. This study evaluates different multi-label algorithms that are discussed in later sections showing their performances.

In the DSS execution phase, the aim is to determine the classes of new documents that were not evaluated by the system during the creation phase. During this phase, the same initial operations of pre-processing and feature extraction of the DSS creation are performed to generate feature vectors.

Finally, the final step on the DSS execution phase is to run the machine learning model, which was built on the previous phase. This model outputs a prediction to aid in deciding the class of a textual document.

After all stage of DSS execution, it is analyzed the feature importance extracted in step (2), with an intent to show wich feature groups are more relevant to the decision model.

3.2. Textual Features

Table 1 lists the features extracted on the second step of the proposed approach and references the works based or inspired on the extraction method, making a total of 29 textual features, being that, nine of then (in bold) used only in feature importance analysis.

The addition of these new features for the feature importance analysis, possibility the divide features extracted into groups with the same scope, where it is possible to make a deep analysis for each type features behavior, what didn't possible with original textual features.

With this, all extracted features were split into five groups: Complexity, Stylistic, Pos tag, Corpus Statistics and Others. Most of them were extracted from the state of the art works, except four features that were proposed, are then: `avgPar`, `missWordC`, `missWordR` and `sumRed`.

3.2.1. Features Complexity

Complexity features are aimed at a complexity capture in a general article. The main idea is based on deep calculations where sentence and word level complexity levels are observed [Horne and Adali 2017].

Table 1. List of extracted features

No	Type	Name	Description	Reference
1	Complexity	avgPar	Average words per paragraph	Proposed
2	Complexity	avgSen	Average words per sentence	[Horne and Adali 2017]
3	**Complexity**	**avgWordSize**	**Average words per size**	**[Lynch and Vogel 2018, Chen et al. 2015]**
4	**Complexity**	**sentences**	**Sentences**	**[Qin et al. 2005]**
5	**Complexity**	**ttr**	**Type-token ratio**	**[Lynch and Vogel 2018, Zhou et al. 2004]**
6	Stylistic	missWordC	Out-of-vocabulary (OOV) words count	Proposed
7	Stylistic	missWordR	Out-of-vocabulary (OOV) words ratio	Proposed
8	**Stylistic**	**upperCase**	**Uppercase letters**	**[Castillo et al. 2013]**
9	**Stylistic**	**quotesCount**	**Quotation marks count**	**[Horne and Adali 2017]**
10	POS tag	ratioADJ	ADJ label frequency	[Shu et al. 2017]
11	POS tag	ratioADP	ADP label frequency	[Shu et al. 2017]
12	POS tag	ratioADV	ADV label frequency	[Shu et al. 2017]
13	POS tag	ratioAUX	AUX label frequency	[Shu et al. 2017]
14	POS tag	ratioCCONJ	CCONJ label frequency	[Shu et al. 2017]
15	POS tag	ratioDET	DET label frequency	[Shu et al. 2017]
16	POS tag	ratioINTJ	INTJ label frequency	[Shu et al. 2017]
17	POS tag	ratioNOUN	NOUN label frequency	[Shu et al. 2017]
18	POS tag	ratioPRON	PRON label frequency	[Shu et al. 2017]
19	POS tag	ratioPROPN	PROPN label frequency	[Shu et al. 2017]
20	POS tag	ratioPUNCT	PUNCT label frequency	[Shu et al. 2017]
21	POS tag	ratioSCONJ	SCONJ label frequency	[Shu et al. 2017]
22	POS tag	ratioSYM	SYM label frequency	[Shu et al. 2017]
23	POS tag	ratioVERB	VERB label frequency	[Shu et al. 2017]
24	**Corpus statics**	**thanking**	**Thanking words**	**[Reganti et al. 2017]**
25	**Corpus statics**	**whQuestions**	**Wh-Questions**	**[Reganti et al. 2017]**
26	**Corpus statics**	**apoWords**	**Apology words**	**[Reganti et al. 2017]**
27	Others	sumRed	Summary reducing rate	Proposed
28	Others	avgSyn	Average synonyms	[Rubin et al. 2016]
29	**Others**	**emotiveness**	**Emotiveness words**	**[Piskorski et al. 2008]**

We propose the usage of average per paragraph (`avgPar`) and per sentence (`avgSen`) grounded in stylometric features [Horne and Adali 2017, Shu et al. 2017]. They are computed tokenizing words of the text and breaking by sentences and line breaks.

With the tokenized words, we extract the average words per size (`avgWordSize`) and the total number of sentences (`sentences`). The type-token ratio (`ttr`) was obtained to capture the lexical diversity of the contained vocabulary in an article [Lynch and Vogel 2018, Zhou et al. 2004]. If the `ttr` value is low, it means that the text has redundancies. Otherwise, it is a text with greater lexical diversity [Dillard and Pfau 2002].

3.2.2. Features Stylistic

Stylistic features are focused on understanding the syntax, grammatical elements and the style of each content and title content in the article. Generally, based on the Natural Language Processing (NLP) [Horne and Adali 2017]

With the tokenized words from the extracted complexity features, we checked against Mac-Morpho [Fonseca et al. 2015], which is a corpus of more than 1 million words in Brazilian Portuguese available on NLTK [Bird et al. 2009], and then counted

every OOV word that is tagged as ADJ, ADV, VERB or NOUN that was not found on the set, assuming it may be an informal word or a neologism.

Then this count is used to extract the total number of OOV words (`missWordC`) and the ratio of OOV words to the total number of tokens (`missWordR`). Also, the count of upper case letters (`upperCase`) features and the quotation marks count (`quotesCount`) were extracted.

3.2.3. Features POS tag

POS tagging terms of the document are a source of several features because each label frequency extracted for the whole document. As shown in [Conroy et al. 2015, Horne and Adali 2017, Shu et al. 2017], POS tags are used as a linguistic descriptor across the fake news detection literature.

3.2.4. Features Corpus Statistics

Corpus statistics features are based on speech act, where actions like apologies, thanking, promises, etc. are taken into account [Reganti et al. 2017, Leech and Weisser 2003]. For this paper, we used just three speech acts: thanking words (`thanking`), wh-questions (`whQuestions`) and apology words (`apoWords`), all adapted to their equivalency in the Portuguese language.

3.2.5. Others

This last group consists of features do not fit in any previous group, and each one of the features has a particular function.

The summary reducing rate (`sumRed`) feature was proposed in this work, taking into account a hypothesis that professional journalists on traditional media vehicles write the lead paragraphs (usually the first paragraph of a journalist text containing the most important information on the text [Bell 1991]) differently from the news written by non-professionals. Thus we generate an automated summary, which is achieved through a variation of TextRank algorithm [Barrios et al. 2016], and compare the result with the size of the original article.

Synonyms are obtained using a pretrained `word2vec` [Mikolov et al. 2013] model by counting the number of most similar terms with a similarity measure higher than a threshold. After that, an average of synonym count (`avgSyn`) is obtained for the document. This feature is related to the semantic validity features proposed in [Rubin et al. 2016], where they consider ambiguity and absurdity of concepts as a characteristic that may be related to satirical texts.

And finally the emotiveness words (`emotiveness`), that is the ratio of modifiers to content words and can be defined as follows [Piskorski et al. 2008].

$$Emotiveness = \frac{\sum adjectives \ \& \ adverbs}{\sum nouns \ \& \ verbs} \quad (1)$$

4. Materials and Methods

4.1. Dataset

For this study, a dataset was created in which Brazilian news was collected from several Brazilian news portals (Brazilian Portuguese). The collecting of data was implemented in two parts: collecting documents from known bias portals and collecting objective fake news from Brazilian checking agencies.

Initially, a web crawler[4] was used to collect a large number of news articles from each website. For each combination of classes, except for the objective-fake news, was chosen portals that have a known purpose. For objective-legitimate news the websites selected were G1[5] and UOL Notícias[6], two of the most visited websites in Brazil according to Alexa ranking[7], filtering by the *politics* tag. Satirical-fake news was collected from Sensacionalista and Diário Pernambucano[8], satirical sites that mimic real news about trending subjects with a humorous aim. For satirical-legitimate news, this study considered sites with bizarre or unexpected events with a jocular tone, such as Surrealista[9], UOL Tabloide[10] and Planeta Bizarro[11].

To validate and collect objective-fake news used in this paper, we used the fact-checking agencies Boatos[12] and Lupa[13] to gather the documents used in the corpus of this study. The fact-checking agencies publish articles verifying truthiness of news articles that are widespread over social networks.

As imbalanced datasets often cause trouble when training the machine learning model [Batista et al. 2004], we selected only those verifications that were checked against textual documents, totalizing 58 documents. The collected objective fake news corpus contains documents from 30 different websites, mostly related to politics and the Brazilian 2018 election. Images and other media were not included in the dataset because they go beyond the scope of this paper.

Lastly, the dataset was built by random sampling from each class combination. The final version had 70 documents that are objective-legitimate, 70 satirical-legitimate, 70 satirical-fake, and 58 objective-fake, as represented in Figure 3. Table 2 show examples of the content of the collected documents.

[4]https://scrapinghub.com/
[5]https://g1.globo.com/
[6]https://noticias.uol.com.br/politica/
[7]https://www.alexa.com/
[8]http://www.diariopernambucano.com.br/
[9]https://www.surrealista.com.br/
[10]https://noticias.uol.com.br/tabloide/
[11]https://g1.globo.com/planeta-bizarro/
[12]https://www.boatos.org/
[13]https://piaui.folha.uol.com.br/lupa/

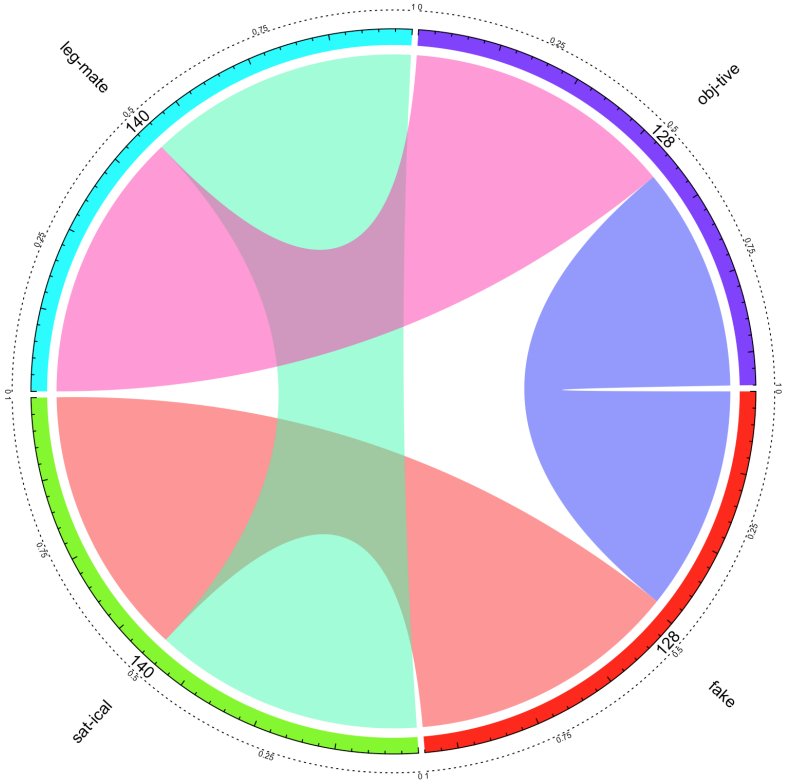

Figure 3. Multi-label circular relation of conceptual classes: Fake, Legitimate (leg-mate), Objective (obj-tive) and Satirical (sat-ical) [de Morais et al. 2019]

The dataset is available at GitHub[14] and contains a total of 268 labeled documents, with an average of 370 tokens per document and a standard deviation of 296, where the smallest instance has 36 tokens, and the largest has 1805.

4.2. Machine Learning Decision

The methods compared used in our DSS are based on multi-label problem transformation, multi-label algorithm adaptation, and multi-class classification algorithms. For the first two, the same multi-class algorithms were explored as the set of base classifiers.

The ML algorithms used as base classifiers and multi-class classification were: Random Forest (RF) [Breiman 2001], Support Vector Machine (SVM) [Cortes and Vapnik 1995] and *k*-Nearest Neighbors (KNN) [Aha et al. 1991], which are grounded in different bias and ML branches.

There exists a wide range of multi-label algorithms [Sorower 2010], but in this study, we focused on evaluating representatives from problem transformations methods and algorithm adaptation for multi-label problems. The problem transformation techniques used on experiments were *Binary Relevance* (BR) and *Label Powerset* (LP) [Zhang and Zhou 2014], so we can evaluate the multi-label approach across different methods.

[14]https://github.com/hugoabonizio/fake-news-multilabel

Table 2. Examples of news content of all conceptual classes

Conceptual Classes		Content
Objective	Legitimate	TSE apresenta previsão do tempo de propaganda no rádio e na TV para cada candidato à Presidência O Tribunal Superior Eleitoral (TSE) apresentou nesta quinta-feira (23) o tempo previsto para a propaganda no rádio e na televisão de cada um dos 13 candidatos à Presidência da República, para a campanha do primeiro turno das eleições deste ano. (...)
Objective	Fake	MST promete guerra civil em caso de prisão de Lula À medida que cresce a força de Lula no seio do eleitorado brasileiro cresce, também, a perseguição movida contra ele pela Operação Lava-Jato e pela mídia golpista. (...)
Satirical	Legitimate	Assaltantes perdem dinheiro de roubo após rajada de vento "Dinheiro na mão é vendaval" é uma grande mentira? Neste caso, um vendaval tirou o dinheiro da mão de bandidos que assaltaram uma agência de viagens em Droylsden, na região da Grande Manchester, na Inglaterra. (...)
Satirical	Fake	Após fim de supletivo em Economia, Bolsonaro dará aulas na UFRJ Após contratar Adolfo Salsisa, professor de economia básica para supletivo dos políticos do DEM, Bolsonaro já tem indicação da Escola Sem Partido para lecionar no Instituto de Economia da UFRJ. Apesar das queixas do professor acerca dos cochilos do aprendiz, Salsisa prevê um futuro presidente bastante graduado em Economia, quiçá mais preparado que Ciro Gomes. (...)

Binary Relevance method decomposes the q labels on q independent binary classifiers that predict whether an instance has a correspondent label [Zhang and Zhou 2014]. This method has a disadvantage when the labels have correlations, but it is not the case in this work because we are classifying into two independent conceptual classes. Unlike Binary Relevance, the Label Powerset method converts each unique label combination into a single-class. It then creates an ensemble where each component targets a random subset of the problem, which addresses the BR's drawback of not considering label correlations.

The combination of ML algorithms as base classifiers and multi-label methods are indicated in this work as BR_KNN, BR_RF, BR_SVM, LP_KNN, LP_RF, and LP_SVM. For the algorithm adaptation methods, we used the ML-*k*NN [Zhang and Zhou 2005], which is an adaptation of the *k*NN algorithm for multi-label data.

Lastly, the final step is to train the models and retrieve their performances over the test set, which is done through a process of stratified cross-validation [Kohavi et al. 1995]. Also, each performance metric was computed after ten iterations on 5-fold cross-validation.

4.3. Metrics

Aiming to evaluate the proposed approach, we tested a set of models and base learner algorithms and compared their results. The metrics used in this comparison are accuracy and f1-score, which are available for multi-label and multi-class classifications.

To define accuracy for multi-label classification, let D be a multi-label evaluation set, Y be the true set of labels, and Z be the predicted set of labels, [Tsoumakas and Katakis 2007] define accuracy as:

$$Accuracy = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \bigcap Z_i|}{|Y_i \bigcup Z_i|} \tag{2}$$

For f1-score we first need to define precision and recall metrics, following definitions by [Tsoumakas and Katakis 2007]:

$$Precision = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \bigcap Z_i|}{|Z_i|} \tag{3}$$

$$Recall = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \bigcap Z_i|}{|Y_i|} \tag{4}$$

For multi-class metrics, having true positive (TP), true negative (TN), false positive (FP) and false negative (FN) classification results, the authors [Olson and Delen 2008] and [Fawcett 2006] defines accuracy, precision and recall as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

Using precision and recall metrics, for both multi-label and multi-class, f1-score is defined by:

$$F1 = \frac{2 * Precision * \ Recall}{Precision + Recall} \tag{8}$$

5. Results

The experiments present in this study have two outcomes: the model's classification performance and insights that can be extracted from it. For model performance, accuracy and f1-score were used, as shown in Figure 4.

As shown in Figure 4, the accuracy of Random Forest and Label Powerset with Random Forest as base classifier are the two highest scores (71% and 70%, respectively) with a difference between them being statistically irrelevant.

That indicates that Random Forest was the machine learning algorithm with the best result for both multi-class and multi-label approaches, appearing in the third place with Binary Relevance as its base classifier. That is a reasonable result because RF is an ensemble approach that creates random weak classifiers, which in turn vote for

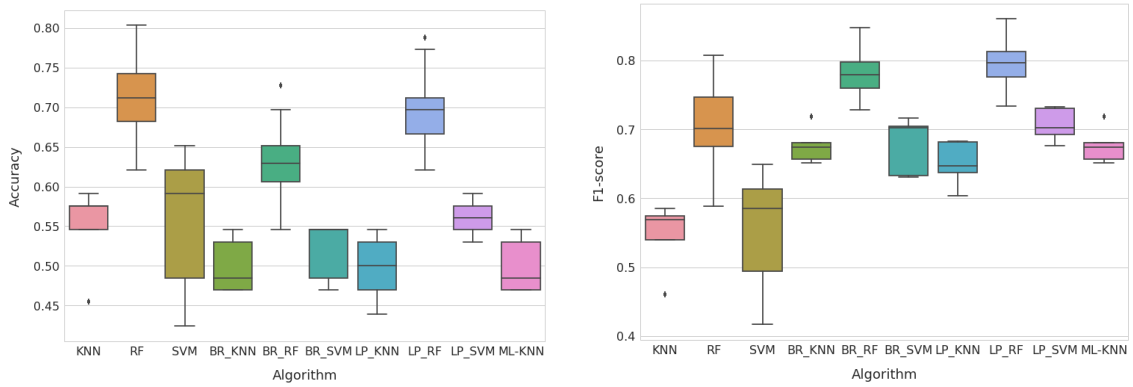

Figure 4. Results of cross-validation through different algorithms using accuracy and f1-score metrics. [de Morais et al. 2019]

the final decision, making the model avoid overfitting and robust to outliers and noise [Breiman 2001].

Concerning on f1-score, the difference between multi-class and multi-label approaches is more significant, with simple RF getting 0.71 and LP_RF achieving 0.80. The two highest f1-scores are from Label Powerset (0.80) and Binary Relevance (0.78) multi-label methods using Random Forest as a base classifier, followed by plain Random Forest in third place.

However, with that result, we can state that the multi-label approach proposed in this work, either by using Label Powerset or Binary Relevance problem transformation methods, is suitable for the problem. On the accuracy measure performance, multi-label methods tied with the best multi-class and f1-score showed multi-label surpassing by a significant amount the classical methods.

The performance of SVM and KNN algorithm combinations was significantly worse than RF counterparts. Still, the plain multi-classes versions were the ones that demonstrated the worse performance, either using accuracy and f1-score. Given the deterministic nature of SVM and KNN, the boxes are generally shorter because they have a lower variance. Except for the plain SVM case, where the box shows a high variance, and it indicates that the hyperparameters of the model have to be tuned to achieve a better result. The nondeterministic nature of RF explains the high variance on the results, which follows an approximately normal distribution.

The outliers appearing on the score results indicate that there are dataset splits made during the cross-validation that was easier or harder for some models to learn the patterns. These outliers can be avoided on future researches using a bigger dataset.

ML-KNN model's performance on f1-score demonstrated that this adapted algorithm, even though better than multi-class approaches, still carries the limitations shown by the plain KNN when compared to an ensemble method (RF).

An important question we attempt to answer is how vital were the features extracted from the dataset, and how much they impact the decision of a document's class.

To answer this question, we extracted the RF variable importance [Genuer et al. 2010], which gives a ranking of the importance of each predictor variable considering the trees created by the algorithm.

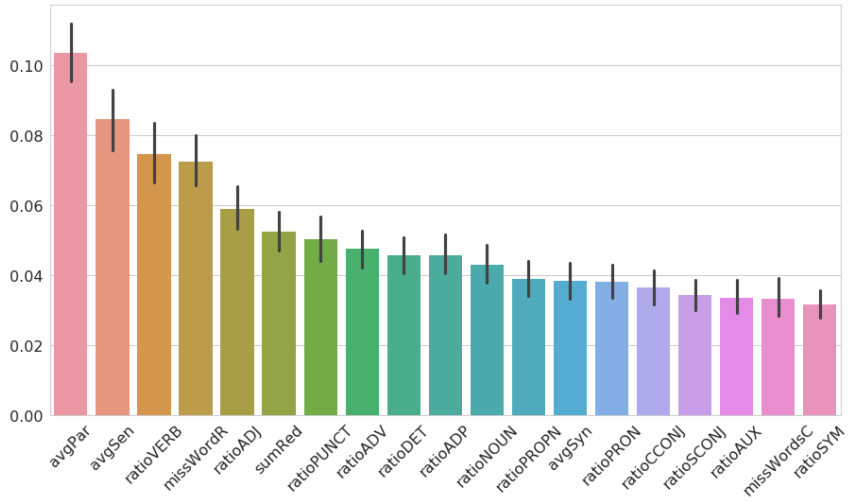

Figure 5. RF importance of textual features explored [de Morais et al. 2019].

Figure 5 presents the ranking of variable importance, where it is possible to notice that the number of words per paragraph and per sentence are the most important variables to describe the dataset, indicating that there is a difference in text size and density that divides the classes. The following features were the frequency of words labeled with VERB tag, the ratio of OOV words, and the ratio of ADJ tag to the total count of words.

The ratio of POS tags showed they were important predictor variables (where the highest rankings were verbs, adjectives, and punctuation some) for classifying fake and satirical news, confirming the results found by [Shu et al. 2017] and [Horne and Adali 2017].

Table 3 presents the average and the standard deviation of the values of the features grouped by label combinations. With these results, it is possible to say that concerning the number of words per paragraph (`avgPar`), there is little difference between objective news and satirical-fake articles. Still, the satirical-legitimate documents have smaller paragraphs and higher variance. The `avgSen` that counts the average words per sentence indicates that objective-fake articles (deceptive news) have substantially smaller sentences, being a sign that this kind of document has a less complex language aiming to be more accessible and superficial.

Concerning those variables, the results demonstrated that objective fake news has a significantly smaller average word per paragraph and more OOV words. The proposed features (`avgPar`, missWordR and `sumRed`) were important descriptors of objectivity and legitimacy of news documents.

Based on the results obtained in the described experiments, we made a detailed analysis of the features importances, where 9 more features were added in addition to the 20 used in the experiments. For this, a multi-label combination of binary relevance with

Table 3. Average value and standard deviation of extracted features grouped by label combinations.

Feature name	Objective Legitimate	Objective Fake	Satirical Fake	Satirical Legitimate
avgPar	41.262 (11.522)	42.602 (13.280)	43.917 (16.467)	28.980 (23.091)
avgSen	21.880 (3.555)	16.804 (3.409)	18.226 (5.820)	20.113 (7.041)
ratioVERB	0.108 (0.019)	0.139 (0.022)	0.136 (0.021)	0.122 (0.029)
missWordR	0.008 (0.005)	0.017 (0.015)	0.017 (0.014)	0.024 (0.042)
ratioADJ	0.045 (0.012)	0.035 (0.019)	0.044 (0.016)	0.047 (0.021)
sumRed	0.284 (0.065)	0.230 (0.093)	0.218 (0.091)	0.233 (0.096)
ratioPUNCT	0.143 (0.026)	0.139 (0.029)	0.122 (0.028)	0.125 (0.040)
ratioADV	0.035 (0.012)	0.042 (0.019)	0.045 (0.021)	0.038 (0.014)
ratioDET	0.096 (0.017)	0.109 (0.016)	0.105 (0.024)	0.106 (0.023)
ratioADP	0.144 (0.023)	0.131 (0.022)	0.134 (0.026)	0.127 (0.029)
ratioNOUN	0.174 (0.030)	0.185 (0.027)	0.174 (0.028)	0.175 (0.034)
ratioPROPN	0.105 (0.039)	0.073 (0.027)	0.103 (0.047)	0.102 (0.074)
avgSyn	6.838 (0.454)	7.031 (0.578)	6.773 (0.598)	6.661 (0.915)
ratioPRON	0.029 (0.012)	0.039 (0.017)	0.034 (0.016)	0.036 (0.017)
ratioCCONJ	0.020 (0.007)	0.021 (0.010)	0.020 (0.011)	0.024 (0.011)
ratioSCONJ	0.012 (0.008)	0.014 (0.009)	0.015 (0.010)	0.013 (0.008)
ratioAUX	0.018 (0.008)	0.022 (0.014)	0.020 (0.014)	0.021 (0.011)
missWordsC	5.271 (4.187)	4.000 (3.522)	3.686 (4.169)	8.293 (16.509)
ratioSYM	0.016 (0.016)	0.012 (0.009)	0.011 (0.010)	0.015 (0.014)

the random forest was chosen, as it was one of the combinations that obtained the best results in the experiments.

However, for this analysis, each performance metric was computed after ten iterations on 10-fold cross-validation instead 5-fold used in the experiments. Moreover, we analyzed each feature group's combination separately in all possible combinations between these feature groups, which can be seen in Figure 6.

As noted, feature groups containing Complexity features have a higher f1-score. Groups containing Stylistic features also have significant performance, following POS tags and Others. The groups containing the Corpus Statistics features did not show to be relevant, which can be seen more clearly when only Corpus Statistics group features are analyzed separately, giving the worst result compared with other groups. Furthermore, it can be observed that the boxplots have a small amplitude which indicates that in most of the compared groups, there was a small variance of values.

To visualize the distribution of samples on the feature space, we applied t-SNE [Maaten and Hinton 2008], a nonlinear dimensionality reduction technique that generates a two-dimensional projection of the dataset. It works by optimizing the Kullback-Leibler divergence between two distribution: a Gaussian probability distribution based on the relationship between each point on the original space and a Student-t distribution that recreates the distribution in the lower-dimensional space. This technique differs from other dimensionality reduction techniques, e.g., Principal Component Analysis, since t-SNE maps complex nonlinear relationships of local and global data structure.

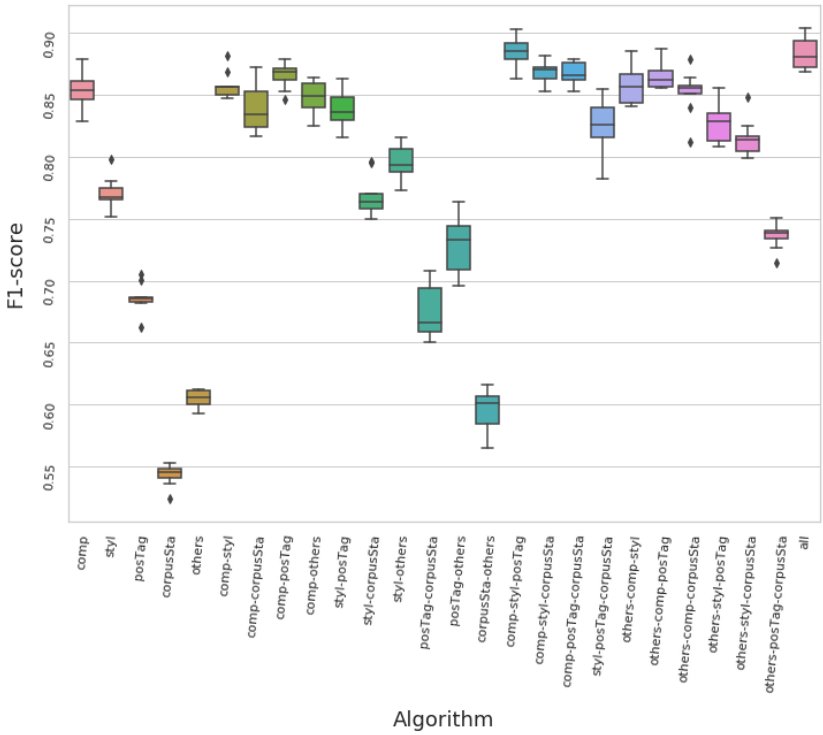

Figure 6. F1-Score of each feature combination.

Figure 7 shows each sample as a point on feature space, highlighting the separability of each conceptual class. With this, it is possible to perceive that the existing separability in the legitimate/fake class shows that there is a predominant side for each of the possibilities. However, even so, the data is dispersed and does not exist an accurate separability between legitimate and fake. The same happens with the objective / satirical class, where it is also possible to see this data separability, showing that the separation between objectivity and satire is a few more homogeneous than in legitimate/fake.

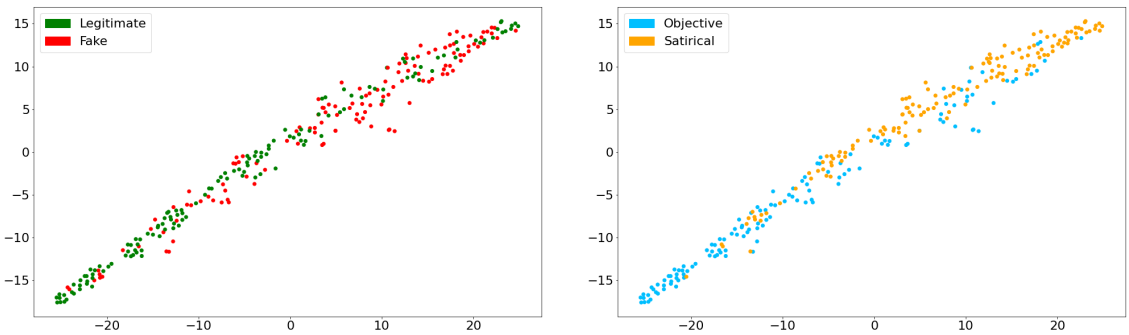

Figure 7. Visualization of samples distribution over the feature space using t-SNE for dimensionality reduction.

The Figures 8 and 9 show the Nemenyi test for f1-score and accuracy, which shows the statistical differences of each one of the analyzed groups. Each one of the connected groups for the horizontal bar is part of the same Critical Distance (CD) bar, i.e. these

interconnected groups are not statically different from each other. As seen in Figure 6, the groups containing the Complexity features obtained a more accurate result.

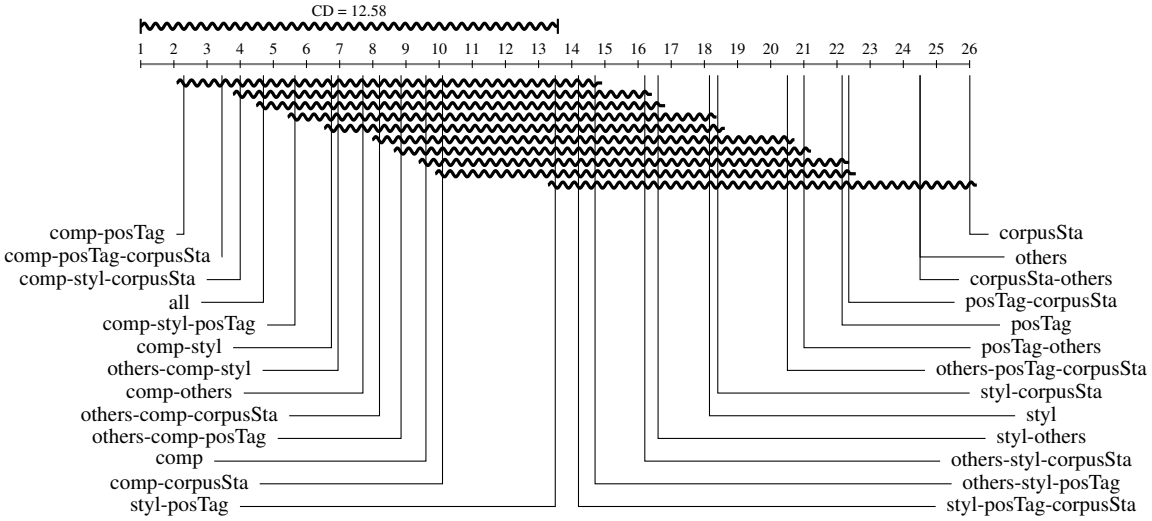

Figure 8. Comparison of models trained with each combination of feature group against each other according to the Nemenyi test, taking into account the accuracy measure. Results from connected groups are not significantly different (at $\alpha = 0.05$).

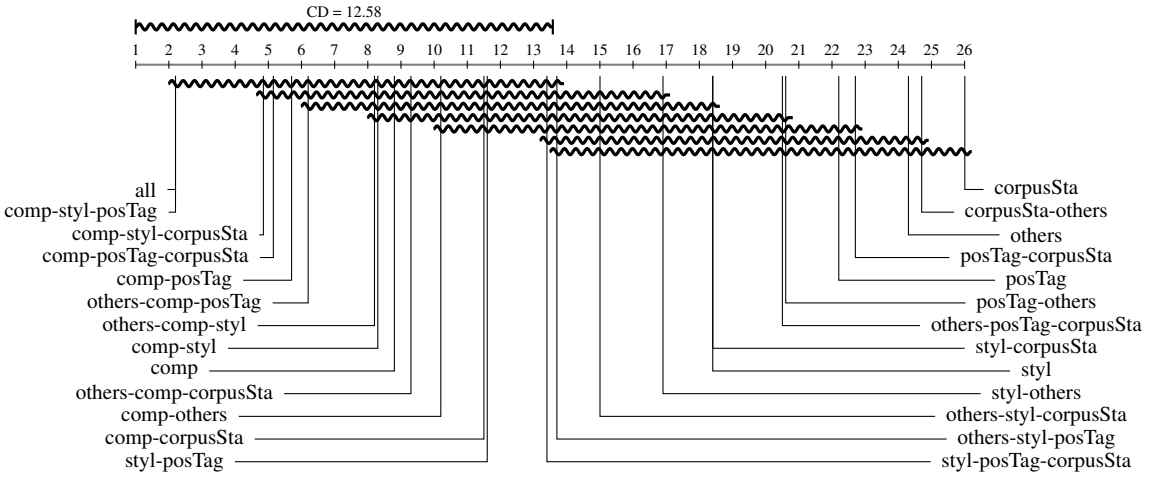

Figure 9. Comparison of models trained with each combination of feature group against each other according to the Nemenyi test, taking into account the f1-score. Results from connected groups are not significantly different (at $\alpha = 0.05$).

In Nemenyi analysis, both accuracy and f1-score can divide feature groups into two large groups, the first containing the groups including the Complexity features that are part of the same CD on the left and the second containing the features Corpus Statistics on the right. Despite this remarkable difference that confirms the larger difference between complexity features and corpus statistics, it is possible to notice that the CD has scales,

where some that that belong to the two large groups have statistical levels considering the CD between them.

5.1. Open Issues

Although the results showed the multi-label approach in our DSS is adequate for the study problem, it is essential to highlight the research only addresses Brazilian Portuguese idiom. However, the results are promising and offer new perspectives to future research in different idioms. It is important to note the chosen features are language independent and can be used in multilingual problems respecting the particularities and possible adaptations for several idioms

Another limiting factor was that our study did not test the performance of our proposed DSS in another domain. We focused on political theme, in addition to using only our results as a parameter for the feature importance analysis process. However, it is important to note that the chosen features are language independent and can be used in multilingual problems respecting the particularities and possible adaptations for each domain.

Feature importance analysis showed which features may be the most relevant to the problem addressed in this study, but paves the way for future analyzes on feature groups in different datasets, idioms, and domains.

6. Conclusions

In this research, we proposed a pipeline for text mining with a classification of news throughout objective/satirical and legitimate/fake conceptual classes. For this, we introduced and used a Decision Support System of news legitimacy and explored a realistic scenario based on a real-life dataset collected from different sources of news. Also, was proposed the usage of a multi-label approach tackling the challenge of classifying four combinations of classes: objective-legitimate, objective-fake, satirical-legitimate, and satirical-fake.

Furthermore, it was used 20 textual features for classification. Also, among these 20 features, four novel features were proposed to improve predictive performance. The ML algorithm, with the best result was RF, which obtained a good result in both multi-class and multi-label approaches. The best performance (f1-score) was achieved by multi-label approaches with the two highest scores being derived from the LP (0.80) and BR (0.78) prevailing over best multi-class (0.71).

Finally, nine textual features were added, and all 29 features were separated into five groups where the extracted features groups were analyzed separately and combine with other groups aiming to help to understand how each feature group can influence at the experiments. The results showed that the complexity features group has a significant influence on the results. Furthermore, the corpus statistics features group showed that it's not relevant to the results, being with the worst performance compared to others, which may symbolize that all the features in this group were not relevant to the results achieved in this paper.

In future research, we intend to increase the number of news in our dataset and compare these results with those obtained this dataset to find out if the pattern seen in the distribution of samples on feature space shown in t-SNE and features importance analyzed tend to keep a pattern.

7. Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Coordination for the National Council for Scientific and Technological Development (CNPq) of Brazil - Grant of Project 420562/2018-4 and Fundação Araucária (Paraná, Brazil).

References

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine learning*, 6(1):37–66.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Almeida, A. M., Cerri, R., Paraiso, E. C., Mantovani, R. G., and Junior, S. B. (2018). Applying multi-label techniques in emotion identification of short texts. *Neurocomputing*, 320:35–46.

Barrios, F., López, F., Argerich, L., and Wachenchauzer, R. (2016). Variations of the similarity function of textrank for automated summarization. *arXiv preprint arXiv:1602.03606*.

Batista, G. E., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29.

Bell, A. (1991). *The language of news media*. Blackwell Oxford.

Bhowmick, P. K. (2009). Reader perspective emotion analysis in text through ensemble based multi-label classification framework. *Computer and Information Science*, 2(4):64.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Castillo, C., Mendoza, M., and Poblete, B. (2013). Predicting information credibility in time-sensitive social media. *Internet Research*, 23(5):560–588.

Chen, Y., Conroy, N. J., and Rubin, V. L. (2015). Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 15–19. ACM.

Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics.

Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, page 82. American Society for Information Science.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

de Morais, J. I., Abonizio, H. Q., Tavares, G. M., da Fonseca, A. A., and Barbon Jr, S. (2019). Deciding among fake, satirical, objective and legitimate news: A multi-label classification system. In *Proceedings of the XV Brazilian Symposium on Information Systems*, page 22. ACM.

Dillard, J. P. and Pfau, M. (2002). *The persuasion handbook: Developments in theory and practice*. Sage Publications.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Fonseca, E. R., Rosa, J. L. G., and Aluísio, S. M. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, 21(1):2.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236.

González-Ibánez, R., Muresan, S., and Wacholder, N. (2011). Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, pages 581–586. Association for Computational Linguistics.

Horne, B. D. and Adali, S. (2017). This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv preprint arXiv:1703.09398*.

Igawa, R. A., Kido, G. S., Seixas, J. L., and Barbon, S. (2014). Adaptive distribution of vocabulary frequencies: A novel estimation suitable for social media corpus. In *Intelligent Systems (BRACIS), 2014 Brazilian Conference on*, pages 282–287. IEEE.

Ishita, E., Oard, D. W., Fleischmann, K. R., Cheng, A.-S., and Templeton, T. C. (2010). Investigating multi-label classification for human values. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Kress, G. (2003). *Literacy in the new media age*. Routledge.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Leech, G. and Weisser, M. (2003). Generic speech act annotation for task-oriented dialogues. In *Proceedings of the corpus linguistics 2003 conference*, volume 16. Lancaster: Lancaster University.

Li, X., Xie, H., Rao, Y., Chen, Y., Liu, X., Huang, H., and Wang, F. L. (2016). Weighted multi-label classification model for sentiment analysis of online news. In *Big Data and Smart Computing (BigComp), 2016 International Conference on*, pages 215–222. IEEE.

Lynch, G. and Vogel, C. (2018). The translator's visibility: Detecting translatorial fingerprints in contemporaneous parallel translations. *Computer Speech & Language*, 52:79 – 104.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

McCombs, M. E. and Shaw, D. L. (1972). The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Olson, D. L. and Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Piskorski, J., Sydow, M., and Weiss, D. (2008). Exploring linguistic features for web spam detection: a preliminary study. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 25–28. ACM.

Poria, S., Cambria, E., Hazarika, D., and Vij, P. (2016). A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*.

Qin, T., Burgoon, J. K., Blair, J. P., and Nunamaker, J. F. (2005). Modality effects in deception detection and applications in automatic-deception-detection. In *Proceedings of the 38th annual Hawaii international conference on system sciences*, pages 23b–23b. IEEE.

Reganti, A., Maheshwari, T., Das, A., and Cambria, E. (2017). Open secrets and wrong rights: automatic satire detection in english text. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 291–294. ACM.

Rubin, V., Conroy, N., Chen, Y., and Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17.

Ruchansky, N., Seo, S., and Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806. ACM.

Saif, H., Fernández, M., He, Y., and Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter. *Ninth International Conference on Language Resources and Evaluation*, pages 810—-817.

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., and Menczer, F. (2017). The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*.

Shoemaker, P. J. and Reese, S. D. (2013). *Mediating the message in the 21st century: A media sociology perspective*. Routledge.

Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. (2019a). defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 395–405, New York, NY, USA. ACM.

Shu, K., Mahudeswaran, D., and Liu, H. (2019b). Fakenewstracker: a tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory*, 25(1):60–71.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Singhania, S., Fernandez, N., and Rao, S. (2017). 3han: A deep neural network for fake news detection. In *International Conference on Neural Information Processing*, pages 572–581. Springer.

Sorower, M. S. (2010). A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18.

Tayal, D. K., Yadav, S., Gupta, K., Rajput, B., and Kumari, K. (2014). Polarity detection of sarcastic political tweets. In *Computing for Sustainable Global Development (INDIACom), 2014 International Conference on*, pages 625–628. IEEE.

Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Zhang, M.-L. and Zhou, Z.-H. (2005). A k-nearest neighbor based algorithm for multi-label classification. In *Granular Computing, 2005 IEEE International Conference on*, volume 2, pages 718–721. IEEE.

Zhang, M.-L. and Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., and Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation*, 13(1):81–106.