Deciding among Fake, Satirical, Objective and Legitimate news: A multi-label classification system

Janaína Ignácio de Morais
State University of Londrina (UEL)
Londrina, Brazil
janainam@uel.br

Hugo Queiroz Abonizio
State University of Londrina (UEL)
Londrina, Brazil
hugo.abonizio@gmail.com

Gabriel Marques Tavares
State University of Londrina (UEL)
Londrina, Brazil
gtavares@uel.br

André Azevedo da Fonseca
State University of Londrina (UEL)
Londrina, Brazil
andre.azevedo@uel.br

Sylvio Barbon Jr.
State University of Londrina (UEL)
Londrina, Brazil
barbon@uel.br

ABSTRACT

Currently, the widespread of fake news has raised on the political class and society members in general, increasing concerns about the potential of misinformation that can be propagated, appearing on the center of the debate about election results around the world. On the other hand, satirical news has an entertaining purpose and are mistakenly put on the same boat of objective fake news. In this work, we address the differences between objectivity and legitimacy of news documents, treating each article as having two conceptual classes: objective/satirical and legitimate/fake. Thus, we propose a Decision Support System (DSS) based on a text mining pipeline and a set of novel textual features that uses multi-label methods for classifying news articles on those two domains. For validating the approach, a set of multi-label methods was evaluated with a combination of different base classifiers and then compared to a multi-class approach. Results reported our DSS as proper (0.80 *F1*-score) in addressing the scenario of misleading news from challenging perspective of multi-label modeling, outperforming the multi-class methods (0.71 *F1*-score) over a real-life news dataset collected from several portals of news.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Learning paradigms** → **Supervised learning**;

KEYWORDS

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

ACM Reference format:
Janaína Ignácio de Morais, Hugo Queiroz Abonizio, Gabriel Marques Tavares, André Azevedo da Fonseca, and Sylvio Barbon Jr.. 2019. Deciding among Fake, Satirical, Objective and Legitimate news: A multi-label classification system . In *Proceedings of XV Brazilian Symposium on Information Systems, Aracaju, Brazil, May 20–24, 2019 (SBSI'19),* 8 pages.

SBSI'19, May 20–24, 2019, Aracaju, Brazil

ACM ISBN 978-1-4503-7237-4/19/05…$15.00
https://doi.org/10.1145/3330204.3330231

https://doi.org/10.1145/3330204.3330231

1 INTRODUCTION

Social networks have brought substantial transformations in the way citizens consume, interpret and process the news. During the 20th century, information circulating in mass media was previously selected by editors of press vehicles who, in general, met the classical journalism criteria - such as timeliness, social relevance and truthfulness [1].

However, the credibility of the press was built precisely from the commitment of many media companies to the practices of professional journalism. In this way, readers and spectators have become accustomed to delegating to the press the task of processing the enormous volume of information about reality and presenting it in an assimilable subset of daily news [2].

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. Therefore, instead of professional journalists, algorithms have become the primary agents in charge of selecting and distributing information that arrives individually to consumers [3].

This dynamic has been coupled with the proliferation of amateur sites that gain high profitability with online traffic, thanks to 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 of users, a multitude of websites is dedicated to producing content of easy and quick viralization, without any compromises with issues unrelated to their profitability. As a consequence, users receive and consume a massive volume of information of dubious origin. According to a survey released in September 2018 by Ipsos[1], about 62% of the Brazilian population believe on the fake news, which shows how aggravating this problem is inserted in our society.

Studies in the field of media literacy [4] have been seeking for years to educate the public on the particularities of the messages in the 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. However, to establish the critique of the contents, it is essential to formulate the reflection on the languages of the media. Without

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

such pedagogy, users have fewer resources to discern between the misleading language of a fake news website and the ironic language of a humour site, for example.

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

However, inattentive readers - and without training in media literacy - often are confused when interpreting humorous texts as true stories. Since irony is a sophisticated language feature, the joke is not always obvious. Moreover, the writers' ability to construct the parody in the form of news seeks precisely to extract humour through allegory. The differences are purposely subtle.

Fake news, on the contrary, is better defined as "news articles that are intentionally and verifiable false and could mislead readers"[5]. The most intense international debate on the subject happened mainly after the election of Donald Trump. Moreover, 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.

Besides that, the irony present in the texts has very particular characteristics, where we can visualize negative or opposite feelings in certain affirmations. When it comes to politics, besides irony, we can associate the sentiment analysis with the disappointment of the people against a particular party, where it influences the results of the researches [6].

Although there are several Text Mining (TM) works in the state of the art addressing the issue of whether a news item is false or not based on its textual content, we believe that a news can carry multiple conceptual class for a single news. Dealing with these proposed modeling, an additional challenge pose to traditional Text Mining towards evolving to a multi-label textual classification.

Common single-label classification address the induction of a model from a set of examples associated with a single label l from a set of disjoint labels L, $\mod |L| > 1$. If $\mod |L| = 2$, we have a binary classification problem. Alternatively, while if $\mod |L| > 2$ it is a multi-class classification scenario. Finally, the multi-label classification, examples are related to a set of labels $Y \subseteq L$. As stated, we consider that a news is associated with a set of possible conceptual class $Y \subseteq L$.

In this 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}

Therefore the aim of this study is 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 related to its falsity (fake/legitimate) or its objectivity (objective /satirical).

The secondary contributions of this work are:

(1) Present our real-life multi-labelled news dataset;
(2) Propose new textual features (Out-of-vocabulary features, Summary reducing rate and Average words per paragraph);
(3) Identify the best machine learning algorithm and textual features in our multi-label news scenario;

The rest 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 the model evaluation. Section 5 presents the results and the 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

Researchers have proposed different methods for detecting fake news in recent years. These methods are intended to detect characteristics in shared texts that represent the content is actually false.

Shu et al. [1] proposed a comprehensive analysis of fake news detection in social media, taking into account characteristics such as false 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, and so on. From this, the authors cited possible ways of solving through machine learning techniques, leaving open ways of exploring this problem with data mining.

Allcott et al. [7] proposed the usage of Natural Language Processing (NLP) and classification models to detect fake news based only on its content. For this purpose, only the textual content of the news was considered and processed with the Recurrent neural network (RNN) and the Gated Recurrent Unit (GRU). The result did not fully exploit the dataset, which resulted in a low comprehensive model.

Improving fake news prediction in news stories, Singhania et al. [8] compared a 3-level hierarchical classifier (words, phrases, and headlines). The author's classification proposal concerning different aspects from the model achieved suitable results. However, its contribution is limited to a binary classification.

Online social networks are a common source of fake news, in this scenario, Shao et al. [9] identified potential bots origin of spreading the fake news on Twitter. The proposed tool recognize the dissemination of misleading information by tracking those accounts responsible for the initial spreading of news and some related patterns. An important discussion of this work is the phenomenon of when a news reaches ordinary people who believe in the content and share it to friends and followers, who trusting the person share it again. This phenomenon, on a large scale, compromises the identification of real conceptual class.

Twitter was also the focus of the study about user behaviour proposed by Ruchansky et al. [10]. The authors proposed a model that combines three characteristics (article text, user response and unique user) to predict fake communications. The results contributed to represent users and articles towards identifying the major sources of risk.

Notwithstanding, only pure detection of fake news is a challenging task, since false news is not yet fully understood as observed by Ruchansky et al. [10]. According to Rubin et al. [11], sarcastic and

[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

ironic news can also be a form of fake news, no matter how great an article of news is made intentionally so that the reader knows that something is real, the news can be confused with the true news depending on how it came to a certain belief in the truthfulness of things. Tayal et al. [6] analyzed tweets 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.

González-Ibánez et al. [12] created a search engine for sarcastic tweet content. With the objective of examining 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 human beings won by a very low difference, but the overall accuracy was low due to difficulties of sarcasm classification from both cases.

Following a line of analytical thinking, Poria et al. [13] proposes the use of Convolutional Neural Network (CNN) to extract feelings, emotion and personality in detection of sarcasm from Twitter. The results obtained surpassed the state of the art, however, it is worth mentioning the experimentation was conducted with a single news source.

Considering proposals related text mining, there are some research in the literature that dealt with multi-label classification. [14–16, 16, 17]. A great part of them devoted to sentiment analysis and multiple topic classifications. It is important to mention the contribution of Almeida et al. [17] in comparison of a wide range of techniques delivering valuable insights about the bias of multi-label techniques.

The most of the related work was based on binary classification (true or false), supported by textual and non-textual features, and focused on specific sources, e.g Twitter. On the other hand, 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 addresses the multi conceptual class of a single news, as stated in the presented multi-label definition.

3 PROPOSED APPROACH

The Decision Support System proposed in this work is based on a pipeline for classifying news documents using stylometric features extracted from text. The aim of this approach is to classify documents into 2 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 be split into two parts: the creation of the DSS and the execution of this decision system to obtain a prediction. Figure 1 illustrates the DSS creation steps, where the model is build with data collected previously. Figure 2 refers to the process of executing the created system either on validating set to evaluate the result or on a production environment with new data.

Figure 1 and Figure 2 illustrates the phases of this pipeline.

3.1 Text processing

Referring to DSS creation, the first step (1) is the pre-processing of the raw dataset, where the text cleaning [17], Part-Of-Speech (POS) tagging [18] and stopword removal [19] are performed. During

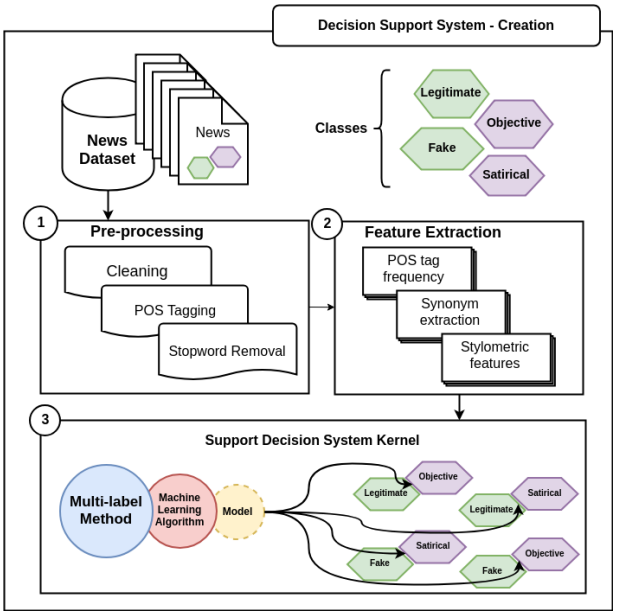

Figure 1: Creation of Decision Support System for detection of fake, legitimate, satirical or objective news.

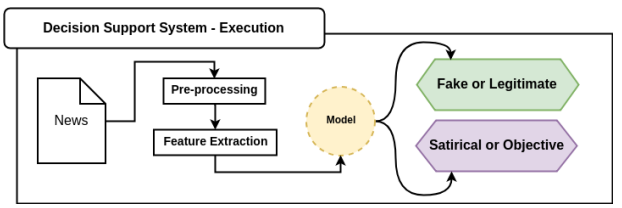

Figure 2: Execution of Decision Support System for detection of fake, legitimate, satirical or objective news.

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 has a high frequency across documents they are considered noise on text data with little discriminative power, and often its removal improve the performance of the model [20], those words are not useful for our purposes.

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

After the processes described above, 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 are used to induce the decision model. After that, on 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 is the kernel of the decision process, where the machine learning algorithm extract patterns to discriminate the classes fashioned by a multi-label domain.

Since we are classifying documents that can be legitimate or fake and satirical or objective, the multi-label approach is more appropriate 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 makes an evaluation of different multi-label algorithms that are discussed in later sections showing their performances.

On 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.

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.

3.2 Textual Features

Table 1 lists the features extracted on second step of the proposed approach and references the works that based or inspired the extraction method, making a total of 20 textual features.

We proposed the usage of average words per paragraph (avgPar) and per sentence (avgSen) grounded in stylometric features [1, 21]. They are computed tokenizing words of the text and breaking by sentences and line breaks. With the tokenized words, we checked against Mac-Morpho[22], which is a corpus of more than 1 million words in Brazilian Portuguese available on NLTK[23], 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).

The `sumRed` feature was proposed in this work taking into account a hypothesis that professional journalists on traditional media vehicles writes the lead paragraphs (usually the first paragraph of a journalist text containing the most important information on the text[24]) in a different way from those news written by non-professionals. Thus we generate an automated summary, which is achieved through a variation of TextRank algorithm[25], and compare the result with the size of the original article.

Synonyms are obtained using a pretrained `word2vec`[26] model by counting the number of most similar terms with a similarity measure greater 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. [11] where they consider ambiguity and absurdity of concepts as a characteristic that may be related with satirical texts.

POS tagging terms of the document are source of several features, each label frequency is generated for the whole document. As shown in [27], [21] and [1], POS tags are used as a linguistic descriptor across the fake news detection literature.

4 MATERIALS AND METHODS

4.1 Dataset

The dataset used in this study was collected through many Brazilian news portals (Brazilian Portuguese). The collecting was implemented in two parts: collecting documents from known bias portals and collecting objective fake news from checking agencies.

For the first part a web crawler[4] was used in order to gather a large amount of news articles from each website. For each combination of classes, except for the objective fake news, portals that have a known purpose were selected. 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-legitimate news was collected from Sensacionalista and Diário Pernambucano[8], satirical sites that mock real news making fun of current subjects. For satirical-legitimate news, this study considered sites that assemble bizarre or unexpected events with a jocular tone, such as Surrealista[9], UOL Tabloide[10] and Planeta Bizarro[11].

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

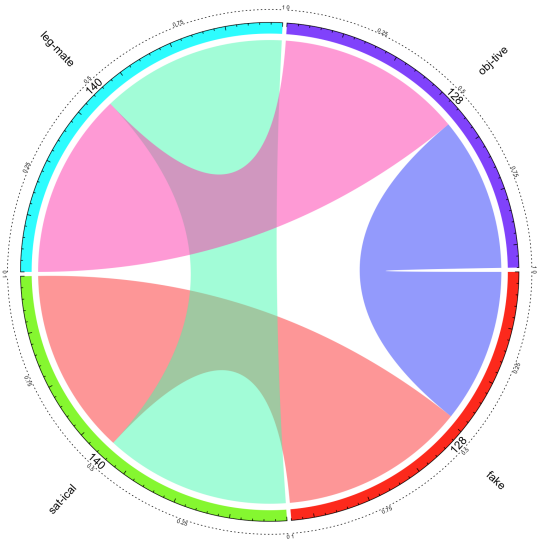

Figure 3: Multi-label circular relation of conceptual classes: Fake, Legitimate (leg-mate), Objective (obj-tive) and Satirical (sat-ical)

[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://piaui.folha.uol.com.br/lupa/
[13] https://www.boatos.org/

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 [21]
3	Stylistic	missWordC	Out-of-vocabulary (OOV) words count	Proposed
4	Stylistic	missWordR	Out-of-vocabulary (OOV) words ratio	Proposed
5	Text Structure	sumRed	Summary reducing rate	Proposed
6	Satirical Cues	avgSyn	Average synonyms	Rubin et al. [11]
7	POS tag	ratioADJ	ADJ label frequency	Shu et al. [1]
8	POS tag	ratioADP	ADP label frequency	Shu et al. [1]
9	POS tag	ratioADV	ADV label frequency	Shu et al. [1]
10	POS tag	ratioAUX	AUX label frequency	Shu et al. [1]
11	POS tag	ratioCCONJ	CCONJ label frequency	Shu et al. [1]
12	POS tag	ratioDET	DET label frequency	Shu et al. [1]
13	POS tag	ratioINTJ	INTJ label frequency	Shu et al. [1]
14	POS tag	ratioNOUN	NOUN label frequency	Shu et al. [1]
15	POS tag	ratioPRON	PRON label frequency	Shu et al. [1]
16	POS tag	ratioPROPN	PROPN label frequency	Shu et al. [1]
17	POS tag	ratioPUNCT	PUNCT label frequency	Shu et al. [1]
18	POS tag	ratioSCONJ	SCONJ label frequency	Shu et al. [1]
19	POS tag	ratioSYM	SYM label frequency	Shu et al. [1]
20	POS tag	ratioVERB	VERB label frequency	Shu et al. [1]

Since imbalanced datasets usually cause trouble when training the machine learning model [28], 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 with politics and Brazilian 2018 election. Image or simple headline news were not included on the corpus because they go beyond the scope of this paper.

Finally, 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 3. Examples of the content of the collected documents are shown on Table 2.

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 to take part 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) [29], Support Vector Machine (SVM) [30] and *k*-Nearest Neighbors (KNN) [31], which are grounded in different bias and ML branches.

There exists a wide range of multi-label algorithms [32], 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) [33], so we can evaluate the multi-label approach across different methods.

Binary Relevance method decomposes the q labels on q independent binary classifiers that predicts whether an instance has a correspondent label [33]. This method has a drawback when the labels have correlations, but it is not the case in this work because we are classifying into 2 independent conceptual classes. The *Label Powerset* method converts each unique label combination into a single-class and 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 referred 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 [34], which is an adaptation of the *k*NN algorithm for multi-label data.

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 [35]. Each performance metric was computed after ten iterations on 5-fold cross-validation.

4.3 Metrics

To evaluate the proposed approach we tested a set of models and base learners algorithms and compared its 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 [36] define accuracy as:

[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. (...)

$$Accuracy = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \bigcap Z_i|}{|Y_i \bigcup Z_i|} \quad (1)$$

For *F1*-score we first need to define precision and recall metrics, following definitions on [36]:

$$Precision = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \bigcap Z_i|}{|Z_i|} \quad (2)$$

$$Recall = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \bigcap Z_i|}{|Y_i|} \quad (3)$$

For multi-class metrics, having true positive (TP), true negative (TN), false positive (FP) and false negative (FN) classification results, [37] and [38] defines:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

Using precision and recall metrics, for both multi-label and multi-class, *F1*-score is defined by:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (7)$$

5 RESULTS

The experiments 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 on Figure 4.

As shown on Figure 4, the accuracy of Random Forest and Label Powerset with Random Forest as base classifier are the two highest

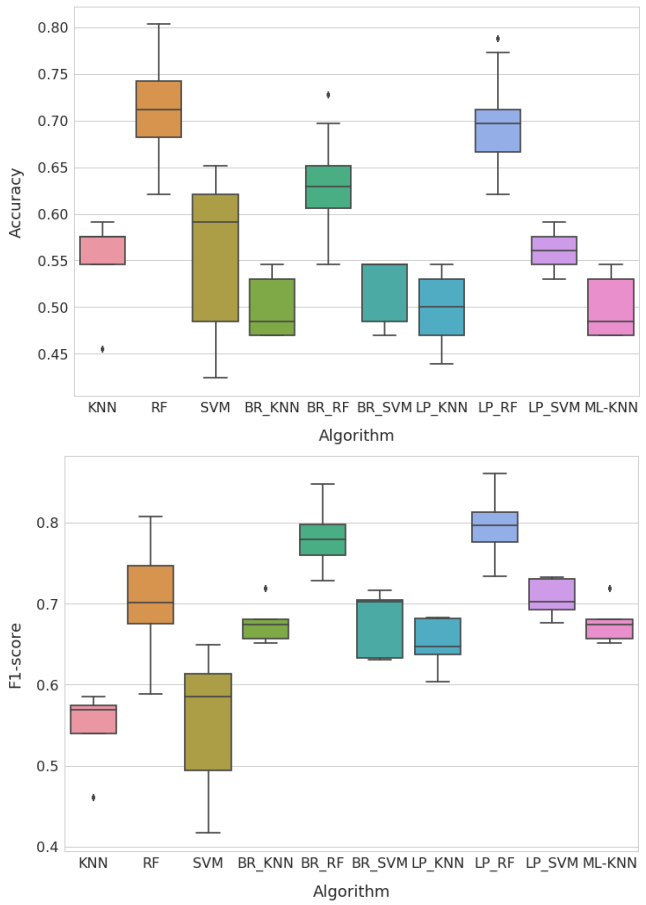

Figure 4: Results of cross-validation through different algorithms using accuracy and *F1*-score metrics.

scores (71% and 70%, respectively) with a difference between them being statistically irrelevant.

This means 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. This is a reasonable result, because RF is an ensemble approach that creates random weak classifiers which in turn vote for the final decision, making the model avoid overfitting and robust to outliers and noise[29].

However, concerning *F1*-score the difference between multi-class and multi-label approaches is more significant, with simple RF getting 0.71 and LP_RF accomplishing 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 base classifier, followed by plain Random Forest on third place.

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, but 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 in order to achieve a better result. The nondeterministic nature of RF explains the high variance on the results, which follows an approximate normal distribution.

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

ML-*k*NN model's performance on *F1*-score demonstrated that this adapted algorithm, even though better than multi-class approaches, still carries the limitations demonstrated by the plain *k*NN when compared to an ensemble method (RF).

An important question we seek to answer is how important were the features extracted from the dataset, and how much they impact on the decision of a document's class. To answer this question we extracted the RF variable importance [39], which gives a ranking of the importance of each predictor variable considering the trees created by the algorithm.

Figure 5 shows the ranking of variable importance, where it is possible to notice that the amount of words per paragraph and per sentence are the most important variables to describe the dataset, indicating that there is a difference on 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 ranking was verbs, adjectives and punctuation some) for classifying fake and satirical news, confirming the results found by Shu et al. [1] and Horne and Adali [21].

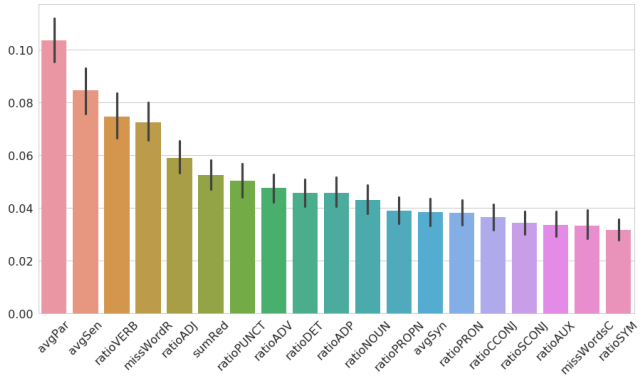

Figure 5: RF importance of textual features explored.

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)

Table 3 shows the average and the standard deviation of the features values grouped by label combinations. It is possible to say from the results that, concerning the number of words per paragraph (`avgPar`), there is little difference between objective news and satirical-fake articles, but the satirical-legitimate documents have clearly smaller paragraphs and a higher variance. The `avgSen`, that counts the average words per sentence, indicates that objective-fake articles (deceptive news) has 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 significant smaller average words per paragraph and more OOV words. The proposed features (`avgPar`, missWordR and `sumRed`) were important descriptors of objectivity and legitimacy of news documents.

6 CONCLUSIONS

In this work, we proposed a Decision Support System to assist the classification of news throughout objective/satirical and legitimate/false conceptual classes. We explored a realist scenario based

on a real-life dataset collected from different sources of news. It was proposed the usage of multi-label approach tackling the challenge of classifying four combinations of classes: objective-legitimate, objective-fake, satirical-legitimate and satirical-fake. Furthermore, four novel textual features were proposed to improve the predictive performance. Based on RF importance, the proposed avgPar overwhelmed the importance of traditional features. The ML algorithm with the best result was RF, which obtained a good result in both multi-class and multi-label approaches. Finally, 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). In future research, we intend to increase the number of news in our dataset and test other idioms.

7 ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

[1] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[2] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[3] Eli Pariser. *The filter bubble: What the Internet is hiding from you.* Penguin UK, 2011.

[4] Gunther Kress. *Literacy in the new media age.* Routledge, 2003.

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

[6] Devendra Kr Tayal, Sumit Yadav, Komal Gupta, Bhawna Rajput, and Kiran Kumari. Polarity detection of sarcastic political tweets. In *Computing for Sustainable Global Development (INDIACom), 2014 International Conference on*, pages 625–628. IEEE, 2014.

[7] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.

[8] Sneha Singhania, Nigel Fernandez, and Shrisha Rao. 3han: A deep neural network for fake news detection. In *International Conference on Neural Information Processing*, pages 572–581. Springer, 2017.

[9] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, 2017.

[10] Natali Ruchansky, Sungyong Seo, and Yan Liu. 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, 2017.

[11] Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 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, 2016.

[12] Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 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, 2011.

[13] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*, 2016.

[14] Emi Ishita, Douglas W Oard, Kenneth R Fleischmann, An-Shou Cheng, and Thomas Clay Templeton. Investigating multi-label classification for human values. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4, 2010.

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

[16] Xin Li, Haoran Xie, Yanghui Rao, Yanjia Chen, Xuebo Liu, Huan Huang, and Fu Lee Wang. 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, 2016.

[17] Alex MG Almeida, Ricardo Cerri, Emerson Cabrera Paraiso, Rafael Gomes Mantovani, and Sylvio Barbon Junior. Applying multi-label techniques in emotion identification of short texts. *Neurocomputing*, 320:35–46, 2018.

[18] Michael Collins. 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, 2002.

[19] Rodrigo Augusto Igawa, Guilherme Sakaji Kido, José Luis Seixas, and Sylvio Barbon. 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, 2014.

[20] Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. *Ninth International Conference on Language Resources and Evaluation*, pages 810—-817, 2014.

[21] Benjamin D Horne and Sibel Adali. 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*, 2017.

[22] Erick R Fonseca, João Luís G Rosa, and Sandra Maria Aluísio. Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, 21(1):2, 2015.

[23] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* "O'Reilly Media, Inc.", 2009.

[24] Allan Bell. *The language of news media.* Blackwell Oxford, 1991.

[25] Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. Variations of the similarity function of textrank for automated summarization. *arXiv preprint arXiv:1602.03606*, 2016.

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[27] Niall J Conroy, Victoria L Rubin, and Yimin Chen. 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, 2015.

[28] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

[29] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[30] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[31] David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.

[32] Mohammad S Sorower. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18, 2010.

[33] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014.

[34] Min-Ling Zhang and Zhi-Hua Zhou. A k-nearest neighbor based algorithm for multi-label classification. In *Granular Computing, 2005 IEEE International Conference on*, volume 2, pages 718–721. IEEE, 2005.

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

[36] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[37] David L Olson and Dursun Delen. *Advanced data mining techniques.* Springer Science & Business Media, 2008.

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

[39] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.