MIND - MAINSTREAM AND INDEPENDENT NEWS DOCUMENTS CORPUS

Danielle Caled, Paula Carvalho, Mário J. Silva
INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal
{dcaled, pcc, mjs}@inesc-id.pt

ABSTRACT

This paper presents and characterizes MIND, a new Portuguese corpus comprised of different types of articles collected from online mainstream and alternative media sources, over a full year period. The articles in the corpus are organized into five collections: facts, opinions, entertainment, satires, and conspiracy theories. Throughout this paper, we explain how the data collection process was conducted, and present a set of linguistic metrics that allow us to perform a preliminary characterization of the texts included in the corpus. Also, we deliver an analysis of the most frequent topics in the corpus, and discuss the main differences and similarities among the collections considered. Finally, we enumerate some tasks and applications that could benefit from this corpus, in particular the ones (in)directly related to misinformation detection. Overall, our contribution of a corpus and initial analysis are designed to support future exploratory news studies, and provide a better insight into misinformation.

***K*eywords** Corpus · Mainstream and Independent sources · Misinformation · Conspiracy

1 Introduction

The present work was developed under the Contrafake project, an interdisciplinary collaboration, involving the academia and official news organizations, with the main goal of creating resources to protect and support communication professionals, citizens, and institutions against misinformation disseminated through digital information sources. As a result of this initiative, we aim to develop machine learning-based tools for (i) creating a set of misinformation indicators; (ii) uncovering information manipulation actions and cyber-attacks; and (iii) early identifying "viral" processes.

The first investigation stage of Contrafake refers to the generation and presentation of misinformation indicators for a given text to information consumers through an online application, named InfoRadar. The purpose of this application is based on the idea that the users themselves should be responsible for selecting the information they want to consume and/or share. Therefore, our task is to provide consumers with a set of relevant indicators that will allow them to verify or to inquire the credibility of the information they are confronted with. In other words, the aim of InfoRadar is not to censor, but to foster media literacy, empowering readers and giving them the means to make informed decisions.

The creation of misinformation indicators was divided into two tasks: (i) the annotation of linguistic and journalistic aspects regarding a set of news articles by communication professionals, which allowed us to identify relevant indicators to distinguish credible news from non-credible news [**OMITTED**], and (ii) the development of a model based on deep-learning techniques to better understand specific characteristics related to a given article.

To train a computational model, we need a representative and diversified collection of textual articles. However, there are few corpora collected for Portuguese, and among the existing ones, none of them are compatible with our needs since (i) they encompass a limited set of articles, with few texts or with little content diversity, and (ii) the existing resources are biased, making the classification task less challenging. Hence, we started by collecting a new corpora, composed of more than 20K articles from 33 Portuguese mainstream and alternative media, aiming at covering different styles, subjects, and communication purposes, over a period of 12 months. The collected articles are organized into five categories, classified as *facts*, *opinions*, *entertainment*, *satires*, and *conspiracy theories*.

We believe this corpus, called MIND (Mainstream and Independent News Documents), will allow us to better understand the main differences among the articles published in the media sources considered in this study, and will facilitate the systematic research on news credibility, allowing the development of computational models to handle disinformation and misinformation.

This paper is organized as follows: We start with a brief related work (Section 2), and then we introduce the data collection comprising our corpus (Section 3). Section 4 presents a set of metrics that will help to perform the linguistic characterization of the classes considered in the corpus. In addition, we present a topic analysis study, which helps to find out which are the main topics (or themes) approached (Section 5). Finally, we discuss the potential use cases of MIND corpus in Section 6, pointing out different tasks and applications. The main conclusions are presented in Section 7.

2 Related Work

Different types of data collections (or *corpora*) have been used in the scope of misinformation detection tasks. For example, we can find corpora composed of short claims, automatically extracted from fact-checking websites (e.g. LIAR [1], PolitiFact [2] and Snopes [3] datasets), while others were crafted using manually created claims based on Wikipedia articles (e.g. FEVER [4]). However, most of the available misinformation corpora rely on the collection of messages posted by users on social media (e.g. Sydney Siege [5], Twitter Series [6, 7, 8], Weibo [7], FacebookHoax [9], BuzzFace [10], FakeNewsNet [11]). Nevertheless, the information included in these types of corpora have specific properties (e.g. size, format, target audience)[1], limiting their usage, particularly if we are interested in exploring more structured, more complex and longer articles.

In terms of content, the data collections most similar to the MIND corpus are the Emergent [12], NELA-GT Series [13, 14, 15] and FacebookHoax [9] datasets. The Emergent collection comprises news articles covering war conflicts, politics and business/technology, and it was created to study how online media handled with unverified information. In a posterior work, the Emergent collection was leveraged into a dataset focusing on veracity estimation and stance classification [16]. NELA-GT Series is a dataset series released annually since 2017; it includes articles in English, collected from different news and media outlets, including mainstream, hyperpartisan, and conspiracy sources. FacebookHoax comprises information related to posts from the Facebook pages related to Italian scientific news and conspiracy websites. Both NELA-GT Series and FacebookHoax contain content associated with mainstream and conspiracy theories sources, assigning the credibility label based on the source-level reliability.

Hardalov et al. [17] assembled four collections with texts written in Bulgarian. These collections consist of (i) credible news (mostly on politics) from a respected newspaper; (ii) articles on general topics from a funny news website; (iii) blog-posts on the political domain classified as "fictitious"; and (iv) credible and fictitious articles made available by the lifestyle section of a television channel's website. This work is related to ours in the sense that the authors resort to similar strategies for building a corpus in a under-resourced language. In addition, both corpora have content collected from mainstream, satirical and fictional sources, covering different topics, such as politics and lifestyle.

Few linguistic resources, however, are available for Portuguese. The first corpus created for the automatic detection of misinformation is Fake.Br [18]. This corpus comprises fake and real news on six main subjects: politics, TV & celebrities, society & daily news, science & technology, economy, and religion. Nonetheless, the Fake.br corpus has two major disadvantages that prevent us from using it in our study. The first one consists on some strong biases (regarding text length, typos and sentiment) [19], which make the analysis simplistic and the classification a less challenging task. On the other hand, this corpus presents a binary collection (with two classes, "fake" or "true"), unsuitable for the diversity of texts we intend to analyze.

3 MIND corpus

The MIND corpus includes 20,278 articles, organized into five different semantic categories, regarding their content and communication purposes: *facts*, *opinions*, *entertainment*, *satire*, and *conspiracy theories*. In this section, we present the methodology used to create this corpus and the corresponding data structure.

[1] Due to restrictions imposed by platforms (for example, a Tweet can contain up to 280 characters), social media messages tend to be shorter, or organized into threads.

3.1 Data collection

The collection period ranges from June 1, 2020 to March 31, 2021, representing a 12-month sample of online content published in Portuguese mainstream and alternative media. Below, we describe the categories considered in the MIND corpus.

3.1.1 Fact. (6000 documents)

Factual articles correspond to news articles collected from nine mainstream Portuguese news websites, addressing national and international topics. These articles were extracted from the following newspapers' sections: *politics*, *society*, *economics*, *technology*, *culture*, and *sports*. Since this information is published by credible news sources, we automatically classify these news stories as facts. This collection is particularly important to study the main properties underlying factual reporting, contrasting it with other types of reporting, including fictional (or satirical) texts and conspiracy theories.

3.1.2 Opinions. (6000 documents)

Opinion articles were collected from the opinion section of ten online mainstream and alternative Portuguese newspapers and magazines. The selection process has prioritized the diversity of authors, even within the same source. This collection is particularly important to distinguish, for example, facts from opinions, which are supposed to involve, respectively, a more objective or subjective language. Regarding subjectivity, the literature shows that it is particularly relevant for distinguishing credible from non-credible information [20, 21].

3.1.3 Entertainment. (6000 documents)

Entertainment articles are the ones published by glossy magazines and tabloids addressing topics such as fashion and beauty, health and well-being, lifestyle, soap operas, celebrity gossip and television shows. This kind of content focus on sensational themes to attract the readers' attention [22], such as divorces, pregnancies, and fights. A common style feature of this type of content is the use of attention-grabbing stories, sensationalism and clickbait strategies, which aren't usually found in factual reporting. To compose this collection, we extracted content published in the celebrity, fashion, beauty, family, lifestyle, and culture sections of six Portuguese magazines, tabloids and newspaper supplements.

3.1.4 Satire. (1029 documents)

This collection of articles gathers material published as fictional or farcical news with a humorous purpose. Satirical content usually involves the use of rhetorical devices, such as irony and sarcasm. In addition to these figures of speech, such texts often explore semantic and pragmatic incongruities [23], which can signal lack of credibility. The articles included in this collection were extracted from two well-known Portuguese websites, self-declared as fictional, humorous, and/or satirical in their editorial guidelines.

3.1.5 Conspiracy Theories. (1249 documents)

Given the difficulty of automatically identifying conspiracy theories, we used a semiautomatic approach to perform the document selection and classification. We explored specific websites that had previously published at least five articles supporting conspiracy theories, particularly about the origin, scale, prevention, diagnosis, and treatment of the COVID-19 pandemic. We focused on the COVID-19 theme as it is a recurring issue, addressed both by the mainstream and alternative media during the collection period. For identifying conspiracy theories, we compiled a list of topics from a selection of conspiracy stories available in Wikipedia[2], and manually inspected a set of candidate websites. This list covers narratives such as the use of the new coronavirus as a biological weapon, the link between COVID-19 and 5G technology, and mRNA-based vaccines altering human DNA. At the end of the inspection, we gathered articles from six different websites that met the selection criteria.

3.2 Data structure

For each article, we collected the (i) metadata available (e.g. article's URL, publication date, authors, category, summary, image and movie URLs embedded in the body of the article), (ii) the article's headline, and (iii) the corresponding

[2] `https://en.wikipedia.org/wiki/COVID-19_misinformation`

Table 1: Examples of headlines* included in MIND.

Col	Examples of headlines
Facts	What is already known about the origin of Covid-19? Women at home quadrupled at school closures
Opinion	Politicians on the stage of the pandemic Should the Left-wing accept that protecting the labour is a taboo?
Entertainment	Drew Barrymore apologizes for being drunk on a TV show Cristiano and Georgina prepare life change
Satire	Reopening of hairdressers and barbers: It is easier to get the vaccine than a haircut Masks in the street: People without glasses forced to guide people with fogged glasses
Conspiracy	China vs. the United States: the oncoming collision Face masks pose serious health risks

* The headlines were translated into English by the authors.

body text. The articles were collected using Python libraries[3] and specific crawlers developed to handle the HTML of unstructured sites.

The full corpus cannot be directly distributed because of copyright held by the sources from which we collected the articles. For this reason, we created an API[4] to download them. The API is provided in the form of a well-documented Python package and a common-separated values file containing the article's metadata. The corpus is referenced using DOI[5] and adheres FAIR Data Principles[6].

The articles from different platforms/domains are stored in different directories, according to its category. In the root directory, we have five different folders referring to the categories considered in the MIND corpus. In a second level of folders, we gather articles grouped according to their source. Inside the source folders, we have the JSON files for each article. Therefore, to access a given article *my_article.json* from a source named *my_source*, classified as *category*, the user needs to access the following path: */category/my_source/my_article.json*.

The article files include all the meta information, headline and body text of the news articles collected using the provided source URLs. Each article file is a JSON object with the following attributes:

authors (*array*): List of authors who have published the article.

tags (*array*): List of article tags assigned by the article's source (e.g., sports, politics, etc).

description (*string*): A short text describing or summarizing the content of the article.

filename (*string*): Name, i.e., identifier, of the file containing the article, automatically generated at the time of its creation.

movies (*array*): List of URLs of all the videos in the article web page (e.g., YouTube, Vimeo, etc).

publish_date (*string*): Date of publication of the article, in the format yyyy-MM-dd HH:mm:ss.

source (*string*): Name of the source which published the article.

top_image (*string*): URL of the best image representing the article.

url (*string*): Article's URL.

headline (*string*): Article's headline.

body_text (*string*): Article's body text.

4 Corpus characterization

Table 1 presents examples of articles' headlines in the MIND corpus and statistics considering the number of sources and documents comprised in each collection. The imbalance among the classes in our corpus highlights the difficulty

[3]The following libraries were used: *newspaper3k* (`https://newspaper.readthedocs.io`); *Beautiful Soup* (`https://beautiful-soup-4.readthedocs.io`), and *feedparser* (`https://feedparser.readthedocs.io`).

[4]To access the corpus, we have published the code implementation at `https://github.com/**hidden-for-blind-review**`. By using this package, users are able to download specific MIND subsets.

[5]`https://doi.org/**hidden-for-blind-review**`

[6]https://www.force11.org/group/fairgroup/fairprinciples

Table 2: 10 most frequent content words (lemmas) in each collection.

	Articles' headline.
Fac	covid-19, new, portugal, year, government, cases, vaccinate, pandemic, death, euro
Opi	new, pandemic, portugal, year, director, letter, time, covid-19, political, country
Ent	new, son, brother, big, covid-19, cristina, ferreira, love, reveal, maria
Sat	marcelo, new, portugal, costa, vaccinate, portuguese, trump, dgs, government, people
Con	covid-19, pandemic, usa, covid, portugal, great, vaccine, vaccinate, new, crisis
	Articles' body.
Fac	year, new, country, person, covid-19, day, pandemic, cases, portugal, president
Opi	year, political, country, new, portugal, time, person, public, day, pandemic
Ent	year, day, marry, son, life, new, person, father, social, family
Sat	know, new, explain, president, portuguese, house, person, vaccinate, day, marcelo
Con	year, person, country, political, great, world, new, social, time, vaccinate

* Terms were translated into English by the authors.

in collecting specific types of data from the Portuguese media, namely satirical and conspiracy content. Particularly regarding satire, there are few known sources with regular and abundant production of such type of content, which could be due to the low profitability, fragile public engagement, or concerns that satire could be perceived as misinformation [24]. On the other hand, the difficulty in obtaining conspiracy theory texts is due to the volatile permanence of these domains on the web (short life-cycle), non-indexed conspiracy pages, which are rarely retrieved by search engines, and the use of other digital platforms (e.g. YouTube and Whatsapp) to disseminate conspiracy narratives [25].

Table 2 presents the top-10 most frequent content words (lemmas) in each collection of documents comprising the corpus. Among the headlines, we can observe that, in general, topics related to the COVID-19 pandemic permeated practically all collections, either directly or indirectly. In our corpus, the most frequent content words include the terms *covid-19* or *covid*. Other related terms (e.g. *cases*, *pandemic*, *vaccinate*, and *vaccine*) are also found, explicitly referring to the new coronavirus. On the contrary, the satirical articles focus mainly on the actors particularly involved in the national and international agendas, such as the Portuguese President (*Marcelo* [Rebelo de Sousa]), the Prime Minister of Portugal ([António] *Costa*), the Directorate-General of Health (Direção-Geral da Saúde; acronym *DGS*), and the former U.S. President (Donald [*Trump*]).

Another characteristic that is common to almost all collections is the use of terms associated with the national political context, such as *government*, *country*, *Portuguese*, and *Portugal*. In contrast to the political character of the remaining collections, entertainment news focus mainly on popular reality shows (*Big Brother*), and entertainers (e.g. *Cristina Ferreira*, who is a famous Portuguese TV personality), personal relationships (*son*), sentiment and emotions (*love*).

As expected, the trend observed in headlines is reinforced in the body. However, entertainment news and satire deviate from the pandemic theme, with the former giving more emphasis on personal relationships and everyday life (e.g. *life*, *family*, *son*, *father*, *marry*), while the latter focuses on the Portuguese presidency (*president*, *Portuguese*, *Marcelo*). Interestingly, although COVID-19 is still widely mentioned in the body of conspiracy articles, there is a higher prevalence of terms such as *world*, pointing out to the global nature of this phenomenon. In addition, it is interesting to stress the prominence of the terms *new* and *year* in almost all the collections, which are probably linked to the emergence of the *new* virus and its impact in the *year* considered in our corpus.

Table 3 presents some statistics on the corpus, derived from of a set of metrics often used in computational linguistics to obtain common textual features that would help characterizing the (mainstream and alternative) news texts included in our data collection. In particular, we considered quantitative metrics related to style and text complexity, such as the length of the headline and body text, which estimate the average number of sentences and words they comprise. We have also calculated the average number of words per sentence, which helps to find potential complex sentences. The results presented in Table 3 show that opinion headlines tend to be shorter, while satire headlines are longer, when compared to the articles from the remaining collections. In contrast, the body text of satirical articles tends to be comparatively quite shorter, both in terms of number of sentences and number of words. This may indicate that the story introduced by the headline is not deeply developed in the body. The most extensive articles belong to the conspiracy theory collection, which are, on average, up to three times longer than the factual news. This contradicts previous studies focused on Portuguese (e.g., [18, 26]), which report false articles as shorter than credible articles. In terms of sentence complexity, factual news tend to be more complex, comprising in average 32 words per sentence, whereas entertainment articles use, in average, simpler linguistic structures.

Table 3: Characterization of MIND corpus in terms of style and complexity (w refers to words, s refers to sentences).

		Fact	Opinion	Entertainment	Satire	Conspiracy
Headline	#s	1.12 +/- 0.34	1.08 +/- 0.31	1.12 +/- 0.33	1.01 +/- 0.15	1.03 +/- 0.18
	#w	11.76 +/- 2.95	7.17 +/- 3.59	12.1 +/- 3.40	14.58 +/- 3.65	10.36 +/- 5.63
	#w/s	10.94 +/- 3.14	6.76 +/- 3.32	11.28 +/- 3.62	14.48 +/- 3.71	10.14 +/- 5.46
Body text	#s	13.72 +/- 15.4	28.31 +/- 19.92	15.61 +/- 15.01	5.29 +/- 6.38	55.49 +/- 63.36
	#w	414.75 +/- 383.29	672.24 +/- 428.78	297.13 +/- 263.06	115.42 +/- 85.03	1372.34 +/- 1510.25
	#w/s	32.63 +/- 29.48	25.53 +/- 8.54	20.92 +/- 9.55	27.09 +/- 11.05	26.05 +/- 7.52

Table 4: Linguistic characterization of MIND corpus.

	Fac	Opi	Ent	Sat	Con
Adjective ratio	0.75	0.70	0.52	0.56	0.71
Adverb ratio	0.70	0.70	0.63	0.75	0.66
Conjunction ratio	0.66	0.63	0.57	0.57	0.60
Noun ratio	0.97	0.92	0.90	0.92	0.92
Number ratio	0.44	0.20	0.22	0.18	0.26
Pronoun ratio	0.49	0.53	0.50	0.54	0.52
Verb ratio	0.98	0.95	0.95	0.95	0.92
Expressivity	0.32	0.40	0.32	0.34	0.38
Modifiers	0.23	0.27	0.23	0.24	0.26
Pausality	4.87	3.58	3.45	3.56	3.55
Redundancy	0.31	0.33	0.30	0.32	0.32
Indefinite pronoun ratio	0.50	0.59	0.44	0.53	0.50
Personal pronoun ratio	0.24	0.35	0.38	0.34	0.35
Modality	0.11	0.13	0.07	0.13	0.12

We have also considered other linguistic attributes specifically related to text credibility [27, 28]. Table 4 summarizes a set of linguistic measures concerning the (i) ratio of occurrence of part-of-speech (PoS) tags in sentences from each collection; (ii) *lexical expressivity* (i.e. the ratio of adjectives and adverbs to the content words they usually modify, respectively nouns and verbs); (iii) the prominence of *modifiers* in text (i.e. the ratio of adjectives and adverbs to the total number of content words); (iv) the *pausality* (i.e. the number of punctuation marks over the total number of sentences); (v) *redundancy* (i.e. the total number of function words over the total number of words); (vi) the ratio of indefinite and personal pronouns (i.e. the total number of indefinite and personal pronouns over the total number of sentences), and the (vii) modality ratio (i.e. total number of modal verbs over the total number sentences)[7].

Table 4 indicates that entertainment and satirical news make use of fewer conjunctions and adjectives, which reinforces the idea that, in such type of content, authors opt for simpler syntactic constructions and have preference for objective language. In addition, entertainment articles use comparatively fewer modal verbs and indefinite pronouns, which attests the idea that they have a direct and focused narrative. On the other hand, factual news make use of more conjunctions and punctuation marks, which may prove the use of a more complex syntax. Furthermore, these articles make use of more numbers, which are typically associated with the news credibility [29], and fewer personal pronouns, which can only be found in the quotations or citations included in the news text. The data shown in Table 4 also suggests that opinion and conspiracy articles are quite similar, with the exception of a slightly more pronounced use of indefinite pronouns in opinion articles.

5 Topic Analysis

We performed a simple topic analysis to understand which are the most prevalent themes in each collection of articles. This analysis was performed through the application of the Latent Dirichlet Allocation (LDA) model, an unsupervised generative probabilistic approach. This approach relies on the idea that documents are random mixtures over latent topics, where a topic is characterized by a probabilistic distribution over terms [30]. We applied LDA MALLET[8]

[7] In this case, we restricted the modality to the most frequent modal verbs in Portuguese (e.g. *poder*, *dever*, *ter de*, *precisar*), which may indicate a possibility, impossibility, contingency, or necessity.

[8] `http://mallet.cs.umass.edu/`

Table 5: Topics extracted from the MIND corpus (with their average proportions across the corpus) and the corresponding topic terms.

No.	Topic name	Topic terms
1	COVID-19 & economy (11%)	pandemic, people, lisbon, millions, euros, europe, politics, economy
2	Soap operas & celebrities (9%)	day, love, life, sic, cristina ferreira, father, video, program
3	TV shows (8%)	home, big brother, daughter, mother, pedro, work, joana, parents
4	Crimes & accidents (8%)	health, year, bank, first, north, sister, moment, minister
5	Second wave & national elections (8%)	christmas, time, future, lack, democracy, fear, portuguese, second
6	Vaccines (7%)	vaccine, portugueses, vaccines, porto, vaccination, ue, children, vacations
7	Sports (7%)	tvi, back, benfica, portuguese, return, first, victory, sporting
8	Emergency state (7%)	state, marcelo, president, death, tap, support, history, emergency
9	Soap operas & politics (7%)	government, maria, country, psd, chega, diary, sex, national
10	U.S. Election (7%)	trump, usa, woman, lockdown, days, man, police, elections
11	COVID-19 cases & deaths (7%)	covid, portugal, cases, deaths, dead, biggest, hospital, infected
12	Crises (7%)	costa, crisis, plan, director, letters, coronavirus, risk, court
13	Celebrity gossip (6%)	son, years, end, died, family, world, party, final

implementation over the articles' headlines comprising the MIND corpus to get a general understanding on the topics present in the entire corpus. The number of topics (n_t) was chosen by optimizing the topic coherence score [31], which estimates the degree of semantic similarity between high scoring terms in the topic. We considered n_t values ranging from 2 to 20. Table 5 presents each of the 13 extracted topics in the MIND corpus. The topics names were manually assigned based on the corpus inspection.

Regarding the topics extracted from the MIND corpus, we realized that the COVID-19 pandemic is - as expected - the most prominent topic, with strong impact on diverse areas, including health, economy, and politics. In fact, references to this theme occur in several identified topics, either explicitly, as in the cases of Topics 1 (*COVID-19 & economy*), 5 (*Second wave & national elections*), 6 (*Vaccines*), 8 (*Emergency state*), 11 (*COVID-19 cases & deaths*), or implicitly, as in Topics 7 (*Sports*), 10 (*U.S. Election*), 12 (*Crises*) and, 13 (*Celebrity gossip*).

We also measured the occurrence of the extracted topics in each collection comprising the MIND corpus (Figure 1). Entertainment is the most dissonant collection regarding the topic distribution. As expected, this collection mostly addresses Topics 2 (*Soap operas & celebrities*), 3 (*TV shows*), and 13 (*Celebrity gossip*). Factual news tend to approach the topics discussed in opinion articles, with the exception of Topic 11, focusing on the COVID-19 new cases and deaths, and Topics 5 and 10, which rely on the analysis of different election scenarios, both at national and international context. There are still many similarities between factual news and conspiracy theories, especially in what concerns the topics on the COVID-19 pandemic. Concerning the conspiracy articles, issues related to vaccines (Topic 6) and to the economic and health impact of COVID-19 (Topics 1 and 11) are widely addressed. Satirical articles have the peculiarity of conjugating unrelated entities, benefiting from the incongruity between them to evoke comic and unexpected situations [23]. This is also observable in our corpus, in which the topics mostly addressed in satirical sources combine, for example, sports (in particular, soccer) and health (in particular, vaccination).

6 Potential use cases of the MIND corpus

The corpus portrays textual content published by the Portuguese media during a one-year period, including a huge variety of articles collected from different sources from the mainstream and alternative media. The articles are classified into different semantic categories, presenting different writing styles, communication purposes, and different levels of credibility. These characteristics make this corpus unique, and it could assist, among others, the following tasks:

- Misinformation detection. Computational models can be trained using the different data collections in MIND corpus as a means to identify credible content. This is in line with previous research [17, 9, 14], which resort to satirical and conspiracy theories sources as a proxy for content (lack of) credibility.
- Identification of authorship at the source level. Currently, the MIND corpus has articles published by 33 different mainstream and alternative media channels.
- Content analysis, taking into consideration the temporal dynamics. In particular, it enables to study how the Portuguese media deals with different topics and entities, in a given period of time. Also, the time frame provided by MIND corpus allows to study the media bias in the Portuguese press, and compare the conventional news with dissonant voices in sources that defy mainstream narratives.

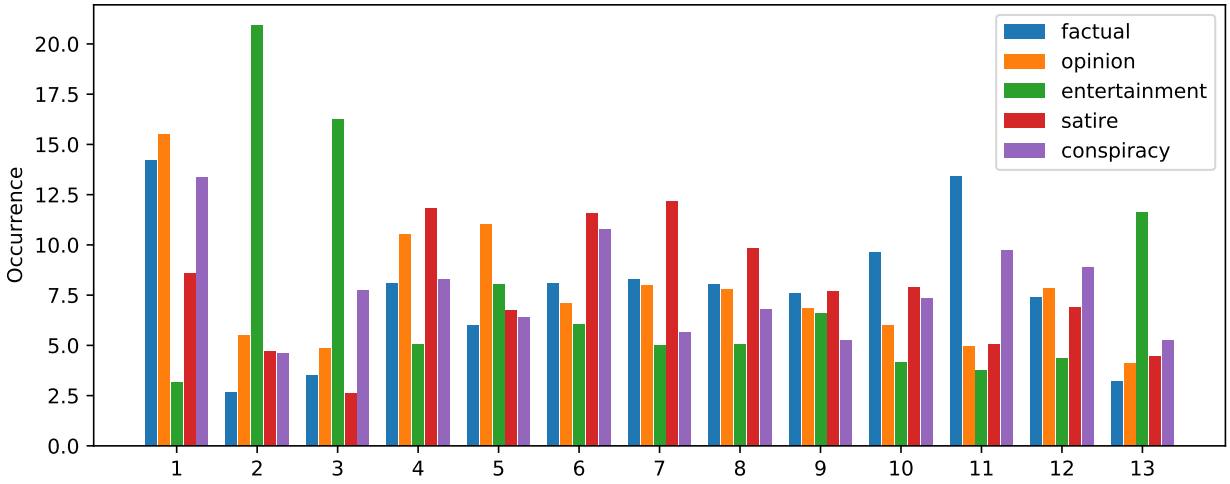

Figure 1: Occurrence of topics in MIND corpus' collections.

- Analysis of clickbait techniques, mainly used in entertainment news, aiming to attract the attention of the reader with sensationalist calls.
- Analysis of sentiment polarity and emotion. It would be interesting to compare the polarity and intensity of sentiment and emotion conveyed in factual news, which are supposed to be neutral towards the subject, to opinionated pieces, which reflect the author's opinion and have no pretensions to be impartial; entertainment news, which frequently address personal affairs; satires, which resort to strategies to make the news funny; and conspiracy theories, which may appeal to reader's feelings (e.g. fear or anger) and action.
- Automatic summarization, by comparing the headline and the corresponding body of each article, taking into consideration the variety of language and pragmatic strategies used in the different collections.
- Issue framing and narrative slant. MIND corpus can also be used to study specific events such as the presidential elections held in Portugal on January 24, 2021, and the serious worsening of the COVID-19 pandemic in Portugal, which peaked in January 2021.

7 Conclusion and Future Work

In this paper, we presented a new corpus of textual documents published by different Portuguese mainstream and alternative sources. The MIND corpus tries to fill a gap in the literature, offering textual material that will enable to perform a variety of studies, from social sciences to computational journalism. In particular, this corpus can help answering research questions involving the study of news credibility, considering a variety of aspects, such as temporal dimension, media source, and communication purpose. It may also support the development of several NLP tasks, including misinformation detection, authorship attribution, and automatic identification of fallacies (for example, based on the conspiracy theories that surround the new coronavirus pandemic). This corpus is still being updated, with the progressive addition of new types (or classes) of documents, extension of the selected sources, and inclusion of new articles. In addition, we intent to create a new collection, addressing especially politically biased sources, observing opposing leanings in the political spectrum.

References

[1] William Yang Wang. "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[2] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium, 2018. Association for Computational Linguistics.

[3] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International*

Conference on World Wide Web Companion, WWW '17 Companion, page 1003–1012, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[4] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[5] Cynthia Andrews, Elodie Fichet, Yuwei Ding, Emma S. Spiro, and Kate Starbird. Keeping up with the tweet-dashians: The impact of 'official' accounts on online rumoring. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 452–465, New York, NY, USA, 2016. Association for Computing Machinery.

[6] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1867–1870, New York, NY, USA, 2015. Association for Computing Machinery.

[7] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 3818–3824. AAAI Press, 2016.

[8] Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[9] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. Some like it hoax: Automated fake news detection in social networks. In *2nd Workshop on Data Science for Social Good, SoGood 2017*, pages 1–15. CEUR-WS, 2017.

[10] Giovanni Santia and Jake Williams. Buzzface: A news veracity dataset with facebook user commentary and egos. volume 12, pages 531–540, Jun. 2018.

[11] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[12] Craig Silverman. Lies, damn lies and viral content. Technical report, Tow Center for Digital Journalism, Columbia University, 2015.

[13] Maurício Gruppi, Benjamin D Horne, and Sibel Adalı. NELA-GT-2019: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2003.08444*, 2020.

[14] Benjamin Horne, Sara Khedr, and Sibel Adali. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. volume 12, Jun. 2018.

[15] Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adalı. NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(1):630–638, Jul. 2019.

[16] William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California, June 2016. Association for Computational Linguistics.

[17] Momchil Hardalov, Ivan Koychev, and Preslav Nakov. In search of credible news. In Christo Dichev and Gennady Agre, editors, *Artificial Intelligence: Methodology, Systems, and Applications*, pages 172–180, Cham, 2016. Springer International Publishing.

[18] Rafael A. Monteiro, Roney L. S. Santos, Thiago A. S. Pardo, Tiago A. de Almeida, Evandro E. S. Ruiz, and Oto A. Vale. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *Computational Processing of the Portuguese Language*, pages 324–334, Cham, 2018. Springer International Publishing.

[19] Renato M. Silva, Roney L.S. Santos, Tiago A. Almeida, and Thiago A.S. Pardo. Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, 146:113199, 2020.

[20] Kristoffer Holt, Tine Ustad Figenschou, and Lena Frischlich. Key dimensions of alternative news media. *Digital Journalism*, 7(7):860–869, 2019.

[21] Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 603–612, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[22] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[23] Paula Carvalho, Bruno Martins, Hugo Rosa, Silvio Amir, Jorge Baptista, and Mário J. Silva. Situational irony in farcical news headlines. In Paulo Quaresma, Renata Vieira, Sandra Aluísio, Helena Moniz, Fernando Batista, and Teresa Gonçalves, editors, *Computational Processing of the Portuguese Language*, pages 65–75, Cham, 2020. Springer International Publishing.

[24] Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly M. Hoffman, Lindsay Huth, Zhenya Ma, Mayanka Jha, Misbah Khan, Varsha Kori, Elo Lewis, George Mirano, William T. Mohn IV, Sean Mussenden, Tammie M. Nelson, Sean Mcwillie, Akshat Pant, Priya Shetye, Rusha Shrestha, Alexandra Steinheimer, Aditya Subramanian, and Gina Visnansky. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 17–21, New York, NY, USA, 2018. Association for Computing Machinery.

[25] Felipe Bonow Soares, Raquel Recuero, Taiane Volcan, Giane Fagundes, and Giéle Sodré. Research note: Bolsonaro's firehose: How Covid-19 disinformation on WhatsApp was used to fight a government political crisis in Brazil. *The Harvard Kennedy School Misinformation Review*, 2021.

[26] Ricardo Moura, Rui Sousa-Silva, and Henrique Lopes Cardoso. Automated fake news detection using computational forensic linguistics. In *Progress in Artificial Intelligence*. Springer International Publishing, 2021.

[27] Lina Zhou, Judee K Burgoon, Douglas P Twitchell, Tiantian Qin, and Jay F Nunamaker Jr. A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, 20(4):139–166, 2004.

[28] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), September 2020.

[29] A. Willem M. Koetsenruijter. Using numbers in news increases story credibility. *Newspaper Research Journal*, 32(2):74–82, 2011.

[30] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, 2019.

[31] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA, 2015. Association for Computing Machinery.