UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Breno de Sousa Matos

Towards Misinformation Span Detection

Belo Horizonte
2024

Breno de Sousa Matos

Towards Misinformation Span Detection

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Rodrygo Luis Teodoro Santos
Co-Advisor: Fabrício Benevenuto de Souza

Belo Horizonte
2024

Matos, Breno de Sousa.

M433t Towards misinformation span detection [recurso eletrônico] / Breno de Sousa Matos. – 2024.

1 recurso online (72 f. il., color.) : pdf.

Orientador: Rodrygo LuisTeodoro Santos
Coorientador: Fabrício Benevenuto de Souza.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f. 52-65.

1. Computação – Teses. 2. Redes sociais on-line – Teses. 3. Mídia social – Teses. 3. Desinformação – Teses. 4. Processamento da linguagem natural (Computação). I. Santos, Rodrygo Luís Teodoro. II. Souza, Fabrício Benevenuto de. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. IV. Título.

CDU 519.6*73(043)

Ficha catalográfica elaborada pela bibliotecária Irenquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais – ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Towards Misinformation Span Detection

BRENO DE SOUSA MATOS

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. RODRYGO LUIS TEODORO SANTOS - Orientador
Departamento de Ciência da Computação - UFMG

PROF. FABRÍCIO BENEVENUTO DE SOUZA - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO
Departamento de Ciência da Computação - UFMG

PROF. FLAVIO VINICIUS DINIZ DE FIGUEIREDO
Departamento de Ciência da Computação - UFMG

PROF SAVVAS ZANNETTOU
Technology Policy and Management - TU Delft

Belo Horizonte, 30 de agosto de 2024.

Para Marcos e Rosângela.

Acknowledgments

Primeiramente, gostaria de agradecer meus pais pelo apoio incondicional, mesmo nem sempre entendendo ou concordando com minhas decisões. Se hoje eu consigo enxergar mais longe e ter as oportunidades que vocês não tiveram, foi devido ao apoio de vocês desde o começo. Agradeço também à minha irmã, que me acompanha desde sempre.

Agradeço também aos meus orientadores Rodrygo e Fabrício, que me deram liberdade criativa para direcionar a dissertação para algo que eu me interesso e valorizo. Mais do que fazer ciência, eles me ajudaram a enxergar o porquê fazê-la. Agradeço também aos professores Marcos, Flávio e Jussara: pesquisadores com quem tive o prazer de colaborar e aprender.

Tenho a sorte de estar cercado por grandes amigos, que me ajudam e que celebram as conquistas comigo. Primeiramente, meus queridos Guilherme, Catrinque, Rafael, Ramon, Leandro, Hideki, Lucas e Barbosa, que me acompanham desde antes do mestrado. São inúmeras as memórias que tenho com eles e, através de muito apoio, conselhos e risadas, sem dúvida contribuíram muito para essa conquista.

Na computação, eu tive o prazer de conhecer diversas pessoas incríveis que fizeram parte do meu caminho dentro do DCC. Especialmente, Chico, Bruno(s), Isadora, Luiza, Caio e Artur, amigos que fiz na computação e que levo pra vida. Sou grato também ao Kaio, Ana, Bernardo, Júnia e Victoria, que estiveram comigo fora do DCC.

Sou grato ao Ian, Gabriel e Aiala, que me acompanham desde antes de me interessar por computação e com quem tenho o prazer de ter uma amizade que sobrevive a qualquer distância.

Gostaria de agradecer também meus amigos do Locus: Mírian, Vitor, Tales, João, Perez, Isadora e Thiago, que foram essenciais nessa jornada, seja colaborando em projetos ou conversando com um cafezinho na mão. Particularmente, agradeço ao Luiz, que traz uma alegria ímpar a qualquer lugar que vai.

Durante a dissertação, pude colaborar com pesquisadores fantásticos fora da UFMG, como o Savvas, que auxiliou no desenvolvimento desta dissertação, sempre com muita presteza a atenção. Agradeço também ao Bob e ao Manoel, que me receberam tão bem na EPFL e me proporcionaram uma experiência que eu jamais pensei que poderia ter, me permitindo expandir meus horizontes e ver que posso e devo almejar mais.

Por fim, um agradecimento mais que especial para Igor, Gustavo e Rennan, que me ajudaram em alguns dos momentos mais desafiadores que já passei e, sem a ajuda deles, essa conquista não seria possível: O Igor é um amigo que todo mundo deveria ter,

mas que poucos têm a sorte de ter na vida. Gustavo, que é meu amigo antes de sermos amigos, me acompanha (e me atura) há uma década. Por fim, o Rennan, para quem posso escrever várias páginas de agradecimentos, e ainda assim não vai ser suficiente.

"E vou viver as coisas novas
Que também são boas
O amor, humor das praças
Cheias de pessoas
Agora eu quero tudo, tudo outra vez"
(Belchior)

Resumo

A desinformação online é um dos problemas mais desafiadores da modernidade, que apresenta consequências severas, incluindo polarização política, ataques à democracia e riscos à saúde pública. A desinformação se manifesta em qualquer plataforma com uma grande base de usuários, incluindo redes sociais e aplicativos de mensagens. Ela permeia todas as formas de mídia e conteúdo, incluindo imagens, texto, áudio e vídeo. Em especial, a desinformação em vídeo representa um desafio multifacetado para os verificadores de fatos, dado a facilidade com que quaisquer indivíduos podem gravar e distribuir vídeos em várias plataformas de compartilhamento de vídeos. Trabalhos anteriores investigaram a detecção de desinformação baseada em vídeo, focando em se um vídeo compartilha desinformação ou não a nível de vídeo. Embora essa abordagem seja útil, ela fornece apenas uma visão limitada e não facilmente interpretável do problema, dado que não fornece um contexto adicional de *quando* a desinformação ocorre dentro dos vídeos e *qual* conteúdo é responsável por tornar o vídeo desinformativo.

Neste trabalho, tentamos preencher essa lacuna de pesquisa propondo uma nova abordagem para a detecção de desinformação em vídeos, focando na identificação da seção dos vídeos que contêm desinformação, uma tarefa que enquadramos como *misinformation span detection*. Apresentamos dois novos conjuntos de dados para esta tarefa, ambos contendo alegações falsas e o momento do vídeo em que elas aparecem. Transcrevemos o áudio de cada vídeo para texto, identificando o segmento do vídeo em que a desinformação aparece, resultando em dois conjuntos de dados com mais de 600 vídeos com mais de 2.400 segmentos contendo alegações verificadas e anotadas. Em seguida, empregamos classificadores construídos com modelos de linguagem de última geração, e nossos resultados mostram que podemos identificar em qual parte de um vídeo há desinformação com uma pontuação F1 de 0,68. Além disso, também apontamos novas direções para a tarefa de misinformation span detection usando *in-context learning*. Esperamos que nosso trabalho possa auxiliar os verificadores de fatos, além do desenvolvimento de ferramentas automatizadas de detecção de desinformação e moderação automática que estejam alinhadas com as necessidades em evolução das plataformas digitais.

Palavras-chave: misinformation; natural language processing.

Abstract

Online misinformation is one of the most challenging modern issues, yielding severe consequences, including political polarization, attacks on democracy, and public health risks. Misinformation manifests in any platform with a large user base, including online social networks and messaging apps. It permeates all media and content forms, including images, text, audio, and video. Distinctly, video-based misinformation represents a multifaceted challenge for fact-checkers, given the ease with which individuals can record and upload videos on various video-sharing platforms. Previous research efforts investigated detecting video-based misinformation, focusing on whether a video shares misinformation or not on a video level. While this approach is useful, it only provides a limited and non-easily interpretable view of the problem given that it does not provide an additional context of *when* misinformation occurs within videos and *what* content (i.e., claims) are responsible for the video's misinformative nature.

In this work, we attempt to bridge this research gap by proposing a novel approach for misinformation detection on videos, focusing on identifying the span of videos that are responsible for the video's misinformation claim, a task we frame as *misinformation span detection*. We present two new datasets for this task, both containing false claims and the video moment in which they appear. We transcribe each video's audio to text, identifying the video segment in which the misinformation claims appear, resulting in two datasets of more than 600 videos with more than 2,300 segments containing annotated fact-checked claims. Then, we employ classifiers built with state-of-the-art language models, and our results show that we can identify in which part of a video there is misinformation with an F1 score of 0.68. Additionally, we also point to new directions for misinformation span detection using in-context learning. We hope our work can assist fact-checkers and the development of automated misinformation detection and robust automatic moderation tools that align with the evolving needs of digital platforms.

Keywords: misinformation; natural language processing.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Social networks and digital platforms are integral to any user's online experience, constituting an essential part of modern life. These platforms enable a wide range of interactions through various applications, becoming essential to people's daily lives. Despite their numerous benefits, such as bridging distances, fostering more effective communication, and enabling marketing strategies, they also bring substantial problems and new challenges to our society, for example, offering a fertile ground for misinformation campaigns.

Misinformation can be defined as pieces of false information that try to appear legitimate by claiming to be real [97, 12, 106], and is one of the most important issues to surface in recent years, affecting modern life in various ways, such as working as an engine for campaigns that promote attacks on democracy [92, 6], political polarization [4] and radicalization [92, 6], and even health-related issues, as evidenced during the COVID-19 pandemic, with the spread of anti-vaccination misinformation [71, 40]. Misinformation's deep impact on modern life is also evidenced by the surge of counter-acting initiatives, such as the International Fact-Checking Network (IFCN), an institution devoted to combating misinformation and widely recognized for its importance, resulting in a nomination for the Nobel Peace Prize in 2021 [110]. Although initiatives to combat misinformation exist, it is still an open, multi-faceted problem.

One of the main challenges in combating misinformation lies in the complexity of digital platform environments and the various forms in which it can arise. For example, misinformation can be launched through websites that appear reliable sources of information but are, in reality, dedicated to disseminating misinformation, often with political motivations [8]. Moreover, misinformation can manifest in different formats, including news pieces, memes, images, and content shared across social networks, specialized groups, and messaging platforms like WhatsApp and Telegram. It spreads through various mediums, encompassing audio [61, 33, 62], video [48, 107, 24, 79, 87, 77, 47, 91], images [38, 78, 94, 46, 58, 93] , and text-based content [42, 41, 109, 73, 60, 75, 25, 59].

The ubiquity of misinformation online motivates our work, which focuses on analyzing and detecting misinformation through state-of-the-art natural language processing techniques. To this end, as further discussed in Section 1.1, we turn to improving misinformation detection, focusing on video-based misinformation, where we leverage the

transcription of the video's audio to locate where misinformation appears. In addition to proposing a novel misinformation detection framework, we also want to provide a proof-of-concept analysis of how state-of-the-art classifiers perform over noisy text.

1.1 Misinformation in Videos

As mentioned in Section 1, misinformation is one of the most challenging problems in our society in the recent years and can take many forms, which offers a great challenge when proposing solutions. Misinformation in video content represents a particularly complex problem due to the massive amount of videos uploaded daily in platforms like YouTube and TikTok. In a single day, YouTube receives a volume of user-generated videos equivalent to 720,000 hours. [1]

Fact-checking agencies cannot keep up with the rapid spread of online misinformation without tools that facilitate journalists to identify content that is worth fact-checking. Additionally, content moderation in videos is a growing concern for platforms such as YouTube and TikTok, especially with novel regulations, such as the that forces the platforms such as the Digital Services Act (DSA) [16], that force those platforms to remove content that is against their terms and also provide transparency about the moderation process.

Thus, this scenario calls for automated detection methods of misinformation in videos. However, unlike detecting if a textual claim or image is fake, detecting misinformation in videos is particularly challenging as one single video can contain hours of speech and become a very laborious task. Despite their undoubted importance, previous research focused on detecting whether a video shares misinformation or not on a video level [107, 48, 44]. While this approach is useful, it only provides a limited and non-easily interpretable view of the problem given that it does not provide an additional context of *when* misinformation occurs within videos and *what* content (i.e., claims) are responsible for the video's misinformation nature.

In this work, we address the problem of *misinformation span detection* in videos, which involves determining the specific segments of a video where misinformation is present. For example, Figure 1.1 depicts a real 55-minute-long video, which was fact-checked by specialists who pointed out 16 misinformative claims (for an illustrative purpose, we marked with red dots the segments in which the false claims are made). Our effort in this work evaluates the feasibility of automatically spotting the segments of the

[1] https://www.globalmediainsight.com/blog/youtube-users-statistics/#stat

Figure 1.1: Example of a fact-checked video with pointers to misinformative segments

videos where these false claims appear. To do it, we used a methodology based on a three-step approach.

First, we gathered two datasets of videos verified by the fact-checking agency Aos Fatos[2], which is part of the International Fact-Checking Network (IFCN) and one of the most prestigious fact-checking agencies in Brazil. Both datasets contain videos and a set of false claims made in the video. The first dataset contains 538 videos featuring Brazil's former president, Jair Bolsonaro, throughout his 4-year term. The second dataset comprises 78 videos containing electoral fraud claims made by voters during the 2022 Brazilian presidential election. Our second step consists of extracting textual transcripts from these videos and annotating the time in which each false claim appears in the video in order to identify which segments of each video contain misinformation. Finally, in our last step, we explored different evaluation scenarios, testing multiple classification approaches using state-of-the-art language models in order to investigate the feasibility of differentiating the segments containing misinformation from those that do not contain them.

Our evaluation results indicate the feasibility of automated misinformation span detection in videos, pointing to valuable directions for developing tools that can assist fact-checkers and moderation in social media platforms.

To the best of our knowledge, we are the first to approach misinformation span detection in videos. We hope our methodology and results offer guidance for future research on the theme and a baseline for comparison. Our results show the feasibility of using automatic detection for this task but also leave space for improvements. We hope our work can inspire future tools to mitigate the misinformation problem in practice. We also propose two publicly available datasets containing 538 and 77 videos, annotated with

[2]https://www.aosfatos.org/

timing in the videos in which 2,355 false claims and 78 false claims occur, respectively. To the best of our knowledge, these are the first datasets of their kind, and we believe they are valuable resources for the research community.

1.2 Dissertation Statement

This thesis states that online misinformation can be further understood and detected by leveraging state-of-the-art natural language processing methods. Online platforms struggle with the diverse and evolving landscape of misinformation, encompassing not only textual content but also other media forms like images, audio, and video. This prompts us to improve automatic misinformation detection, specifically concerning video content, by employing classifiers based on large-language models on the transcriptions of said videos.

1.3 Dissertation Contributions

Our contributions are the following:

- We further develop the task of misinformation detection on videos by formalizing the task of *misinformation span detection*.
- We propose two novel false information datasets with timestamp labels for *misinformation span detection* in videos.
- We define the first baselines for the task, analyzing the problem in multiple settings and providing a thorough analysis.
- We point to possible factors affecting performance, such as the noise level in the videos analyzed, which can guide further efforts in the task.

1.4 Dissertation Outline

Here we present a brief summary of the contents detailed in each chapter of this dissertation:

- **Chapter 2** presents previous works on misinformation and text-based classifiers related to our domain.
- **Chapter 3** presents our results for misinformation span detection on videos
- **Chapter 4** concludes our work, highlighting main findings and future work possibilities.

Chapter 2

Related Work

This chapter discusses previous works on online misinformation. Section 2.1 provides insight into how misinformation is a central and interconnected problem online. Then, Sections 2.2, 2.3, and 2.4 cover studies on the multiple media forms misinformation can encompass on social media, previous studies on language models for misinformation detection, and in-context learning, all of which are vital for Chapter 3.

2.1 Online Misinformation

Misinformation permeates online environments and is often associated with other societal phenomena, such as abusive language. This section discusses the relationship between misinformation and abusive language, one of the main current issues online. Several works have explored the online abusive language phenomenon before, which has been studied under several names such as hate speech [30], online harassment [39], cyberbullying [22], toxicity [29, 5], microaggressions [9, 3], stereotyping [66, 35], unhealthy conversations [2] and others, and we now go over previous works that explore this phenomenon in its many forms. For instance, Mathew et al. [63] have shown that hateful content spreads faster and can reach a broader audience on social networks, in consonance to what Pennycook et al. [74] and Sylvia Chou et al. [96] concluded. Moreover, Zannettou et al. [108] explored news content and found that political and divisive events are more related to hateful commenting, which shows that the use of abusive language online is directly related to political polarization. Kwok and Wang [53] evidenced the difficulty in analyzing and detecting racism online, specifically on Twitter. Hewitt et al. [43], Rodríguez-Sánchez et al. [86], and Fuchs and Schäfer [36] study misogynistic discourse on Twitter and highlight the challenges of working with such data, with Fuchs and Schäfer [36] focusing on instances of misogynistic language against female politicians, showing an increase of hateful expressions against this demographic on Twitter. Additionally, Rodríguez-Sánchez et al. [86] analyze how sexism is expressed in online conversations in

Spanish on Twitter. Clarke and Grieve [20] analyze racist and sexist tweets under linguistic variations users expressed, exposing distinct dynamics between the two. Other works have also emphasized the use of abusive language against religious groups: Chandra et al. [17] present a study on the problem of detection and categorization of antisemitism in online platforms. In contrast, Saha et al. [88] show evidence of discrimination against Muslims in India while also evidencing how users that employ abusive language, namely fear speech, gather a larger following and are more central in online environments, further evidencing the relevance of studying abusive language in online platforms and how individuals can weaponize this form or discourse to gain relevance.

Beyond the analysis of abusive language, mitigation efforts have been proposed over the years, as Fortuna and Nunes [34] evidenced in their comprehensive survey: Caselli et al. [15] propose a new annotation scheme that aims to assess abusive language regarding intention, effect, and the degree of explicitness of the message. Furthermore, Karan and Šnajder [49] assessed abusive language classifiers on diverse datasets from various sources and language types, revealing poor generalization of these models to different domains, highlighting the need for further studies in the field.

Alternatively, other works also explored online misinformation. Vosoughi et al. [104] analyzed the diffusion of true and false news on Twitter from 2006 to 2017, finding that false information spreads further, faster, and more broadly than true news in various categories, with human users playing a more significant role in spreading false information compared to robots. Blankenship [7] also explored the landscape of misinformation on Twitter, examining 14,545,945 tweets produced in response to the Las Vegas shooting[1] and its second anniversary, aiming to determine the extent of public responses affected by information pollution, and to pinpoint the nature and dissemination of misinformation on Twitter and in news coverage. Nan et al. [67] report on the rapid growth of health misinformation research, highlighting its sources, prevalence, characteristics, and impact, ultimately suggesting that while it originates from various sources, especially mass and social media, efforts to mitigate its effects are showing promise in correcting misperceptions. Furthermore, other works [74, 96] agreed that false information spreads faster than genuine content.

Other works on misinformation focus on proposing mitigation solutions, such as Vicario et al. [103], who introduce a framework that uses users' behavior on social media to predict potential targets for misinformation and fake news, effectively identifying fake news. Paschalides et al. [69] introduce Check-It, a web browser plugin designed to efficiently detect fake news by combining various signals. Saxena et al. [90] address the challenge of changing user opinions by identifying a strategic set of users to counteract misinformation, considering users' biases and social interactions, with successful results demonstrated on Facebook and Twitter datasets. Furthermore, Karduni et al. [50] pro-

[1] https://en.wikipedia.org/wiki/2017_Las_Vegas_shooting

poses Verifi2, an educational tool for combating misinformation, as indicated by interviews conducted with experts from various fields.

The works we discussed so far have evidenced how abusive language and misinformation are key online issues, and other recent works have discussed how these issues are connected. Regarding these problems' interplay, several social studies have theorized about their relationship [23, 70, 21]. Accordingly, Giachanou and Rosso [37] endorsed the importance of more quantitative studies on both problems. The authors presented the evaluation process, datasets, and shared tasks related to online misinformation and hateful content. Also, they mention the importance of textual features in detecting such content, which enforces the importance of studying textual patterns. Another remarkable work that explored online abusive language and misinformation dynamics is from Cinelli et al. [19]. In their work, the authors described how users spread offensive content on the YouTube platform and explored its relationship with misinformation-spreading communities. However, the authors focus on online comments written in Italian by YouTube users, which is a narrow sample of such content online. Finally, Matos et al. [64] analyzed the interplay between abusive language and misinformation in news articles' production patterns, focusing on the textual news content; they performed a textual analysis of online news and concluded that false news presents a higher prevalence of abusive language when compared to real news. The found patterns are consistent across datasets, even when they belong to different topics, highlighting the relationship between these issues.

The works on misinformation and abusive language and, ultimately, their relationship show how central misinformation is in the study of online harm, as it is intertwined with other relevant phenomena. We argue that this motivates further efforts in misinformation detection and also grounds the work presented in this thesis.

2.2 Media-specific Misinformation

Social media platforms enabled much faster communication between users and increased the speed of information spread in general. However, this phenomenon also facilitated the spread of online misinformation, prompting platforms and researchers to present solutions to this problem.

Misinformation can take many forms, and media-specific efforts to detect them have been proposed, such as those targeting text posts on social media (e.g., tweets) [42, 41, 109, 73, 60, 75, 25, 59], images [38, 78, 94, 46, 58, 93, 51], and videos [48, 107, 79, 87, 77, 47, 91, 44].

Among all forms of misinformation, video is one of the most challenging due to the

intrinsic difficulty of working with such data type, which usually requires more processing power than, for instance, text, and previous works have proposed mitigation solutions for the issue. Yi Liaw et al. [107] propose a dataset of conspiracy videos on YouTube and a pipeline to detect such videos. However, they perform classification at a video level, not pointing to where the conspiracy claims are made. Hou et al. [44] propose a similar approach for medical videos, also providing a dataset of annotated YouTube videos containing misinformation on prostate cancer, but using an SVM-based classifier for their experiments.

Other works on misinformative videos focus on the platforms where they were uploaded, such as the work proposed by Hussein et al. [47], which highlights the issue of misinformation on videos by auditing YouTube and evidencing how their recommendation systems can induce users to misinformative filter bubbles and grounding the need for more automated tools for misinformation detection on videos.

Additional works focus on manipulated videos: Sabir et al. [87] focuses on deceptive face manipulation on videos, also referred to as deepfakes, a form of misinformation built through synthetically generated media. Similarly, Pu et al. [77] centers on investigating if deepfake detection methods proposed in the literature generalize to real-world deepfakes.

Other studies focus on short videos specifically: Shang et al. [91] investigate misinformative videos about COVID-19 on TikTok, one of the largest video-sharing platforms, by leveraging captions and video components to propose a classification approach. Qi et al. [79] also focuses on short video fake news and builds a dataset by crawling Chinese fact-checking portals, providing a baseline for binary multimodal detection of fake news videos' detection.

Another important work on misinformative videos was presented by Jagtap et al. [48] where the authors propose a framework to classify videos into misinformation and non-misinformation, analyzing 2125 videos containing information about the vaccines controversy, the 9/11 conspiracy, chem-trails, the moon landing, and flat earth. However, like [107, 44, 91, 47, 79], they also focus on binary classification on a video level, lacking an approach that can infer in which part of the video the misinformation appears.

Considering previous works, we propose a new approach to misinformation detection on videos, further discussed in Chapter 3. Specifically, we set ourselves apart from previous works limited to the binary classification of videos containing misinformation. Our work also differs from previous ones that are limited to short videos. In summary, we propose a general approach to misinformation detection that can be used for videos of varying lengths while identifying which section of the video presents misinformative content, a task we frame as misinformation span detection.

2.3 Language Models for Misinformation Detection

Natural language processing has advanced significantly in light of Transformers-based pre-trained models. Those models, such as BERT [28] and GPT [10], allowed the processing of large corpora in an unsupervised fashion to yield contextual and meaning-rich embeddings. This capacity is due to their quadratic attention mechanism [102], which allows for representing a token given all the other tokens in a sentence, leading to better contextualization and text understanding. This mechanism allowed the Transformer architecture to overcome the limitations of older NLP architectures such as LSTMs and CNNs [102]. Therefore, given their contextual text understanding capabilities, Transformers-based language models' performance is currently state-of-the-art for various tasks [28, 10].

Transformers-based models have also aided automatic misinformation detection. Pelrine et al. [73] have shown that simple Transformers-based baselines, such as BERT and RoBERTa reached state-of-the-art performance for misinformation detection on social media posts, for instance, Twitter. Raza and Ding [81] also employed Transformers-based models for misinformation detection by proposing an encode-decoder model, similar to the BART architecture [55], combined with social media features to detect fake news. Praseed et al. [76] also proposed a Transformers-based model for a similar task: their Transformers-based model ensemble improved the effectiveness of Hindi misinformation detection. Moreover, Truică and Apostol [101] provided comprehensive empirical work showing the performance of various Transformers models for fake news detection: in their work, authors show how their proposed model MisRoBÆRTa compares to other Transformers baselines and their performance in different datasets and parameter settings. Overall, recent work endorses the state-of-the-art performance of Transformers models in misinformation detection tasks, motivating us to employ Transformers-based models for misinformation span detection in Chapter 3.

Other architectures have been proposed in recent years, such as LLaMa [99], which is based on the decoder part of the Transformer, adapting its architecture in several components. Additionally, the LLaMa models available to the public are much larger, parameter-wise, than the BERT models. For comparison, the largest BERT has 340 [28] Million parameters, while the largest LLaMa has 65 Billion parameters [99]. The large number of parameters, along with the updates in architecture and extensive amount of training data, led LLaMa's performance to reach the state-of-the-art in various tasks [100]. Similar performance is also seen in related language models, such as GPT [10] and PaLM [18]. Yet, the large amount of parameters demands a higher training cost, which can sometimes be prohibitive.

2.4 In-context Learning

Large language models (LLMs) have surfaced recently, enabling unprecedented performance in multiple natural language processing tasks. Traditionally, to adapt an NLP model for a new problem or dataset, one would need to do multiple rounds of fine-tuning, which is still the case for language models such as BERT or T5. However, LLMs have fostered a new paradigm in Natural Language Processing: In-Context Learning (ICL), which consists of learning through a few examples in the prompt.

Dong et al. [31] state that the "key idea of in-context learning is to learn from analogy." ICL requires demonstrations, which serve as examples in the prompt, and a query question. The demonstrations and the query are concatenated and fed as input to the model for prediction. However, in an ICL setup, no model parameters are updated. Unlike in a traditional supervised learning setting, the demonstrations are expected to be enough for the model to learn the pattern and make the correct prediction.

Since its proposal, ICL has been used in multiple contexts. For instance, Sahu et al. [89] evaluate one sentiment classification (GoEmotions [26]) and three intent classification datasets (BANKING77 [14], HWU64 [56], and CLINC150 [54]), achieving state-of-the-art performance in all tasks using open source LLMs. They also highlight how larger models are needed to take advantage of more demonstrations in the prompt, as smaller models see a plateau or decrease performance as more demonstrations are used.

Min et al. [65] present one of the most comprehensive studies of ICL in different NLP settings, exploring 142 NLP datasets, including question answering, classification, and paraphrase detection, among others. The authors propose a meta-learning approach in which a pre-trained language model is tuned to do ICL on multiple training tasks; this enables a model to effectively learn a new task during inference without needing parameter updates. This new approach outperforms baselines, including ICL (with no meta-training), and, more surprisingly, yields on-par performance with models 8x bigger and fine-tuned on a specific target task, which showcases the effectiveness of ICL.

Although ICL is widely used in the literature, few works tackle its use in misinformation detection. Related to the domain of this work, Liu et al. [57] explore cross-domain misinformation detection using in-context learning. The authors propose RAEmoLLM, a framework that leverages ICL based on affective information to detect misinformation, which removes the cost of fine-tuning LLMs. Authors also perform experiments with zero-shot and few-shot methods that do not incorporate affective information, showing that doing so is an effective addition to the detection process. Although this work sheds light on how to incorporate ICL in misinformation detection, it does not tackle misinformation in videos, and it especially does not tackle tasks similar to misinformation span detection, a gap we bridge in this work.

Chapter 3

Misinformation span detection

3.1 Problem description

As presented in Section 1.1, we now turn to propose mitigation efforts in this Chapter, specifically, misinformation detection on videos, proposing the task of *misinformation span detection*. The objective of this task is the detection of the spans that make a piece of content misinformative. [1] [2] Specifically, we aim to detect whether a piece of content is misinformative and, in particular, which spans of the content are responsible for the content's misinformative nature. Identifying these spans of false claims is paramount as it can assist fact-checkers and social media operators in providing the necessary context (e.g., warning labels) at the exact time of appearance of the false claims.

Although finding mis/disinformation in videos is greatly important, previous work lacks sufficient data for misinformation span detection. In this light, we build two novel datasets for the task: 1) BOL4Y and 2) EI22, further discussed in Sections 3.2 and 3.3.

3.2 BOL4Y dataset

To build our first dataset, henceforth referred to as BOL4Y, we leverage a list of false claims made by Jair Bolsonaro, Brazil's former president. AosFatos,[3] one of Brazil's biggest fact-checking agencies, compiled a list of 6,685 claims through **Bol**sonaro's **4-Year** presidential term.[4] These claims come from multiple sources, such as interviews, written social media posts, and videos that Bolsonaro shared. Each fact check contains

[1] Our task is analogous to the Toxic Spans Detection task presented by Pavlopoulos et al. [72].

[2] As discussed in the next sections, our approach focuses on the transcriptions of the videos. That is, no visual elements are used for detection.

[3] https://aosfatos.org/

[4] https://www.aosfatos.org/todas-as-declara%C3%A7%C3%B5es-de-bolsonaro/

the following data:

- **Claim:** A sentence that summarizes the false claim.
- **Fact-check:** Fact-check produced by AosFatos' journalists.
- **Broad theme:** The theme and broad topic of the claim (e.g., infrastructure, COVID-19 pandemic).
- **Repetition count:** The number of times Bolsonaro made that claim on other occasions, including the dates for each occurrence.
- **Source:** The link to the source (e.g., social media post) that includes the false claim. Although most claims have repetitions throughout Bolsonaro's presidency, AosFatos only lists the source for one of those occurrences. Also, it includes the category of the source (e.g., interview, live stream, etc.).
- **Media repercussion:** Links to other media websites that published a news piece about the claim.

We created our dataset by scraping AosFatos' website in March 2023, collecting data for 6,685 claims from 1,595 unique sources, which vary and include, for example, news pieces from major outlets, posts on social media, and official declarations on governmental websites. Then, we specifically focused on claims with video-based sources, primarily from social media platforms like YouTube, Facebook, TikTok, and occasionally from news websites. Then, we visited the sources and downloaded the videos, obtaining a set of 525 videos. Also, we note that for 121 claims, AosFatos did not provide a link to the source. However, they provide the transcript of the video that comes from AosFatos' transcription service, Escriba. [5] We complement this dataset with these readily available textual transcripts. Overall, this dataset includes 525 videos sharing false claims and 121 textual transcripts (corresponding to videos sharing false claims) obtained from AosFatos' transcription service. The next subsection details how we built the BOL4Y dataset using the data mentioned. We also include a more in-depth analysis of these videos' metadata, including comments, in Appendix A.

3.2.1 Building BOL4Y

Our methodology for building BOL4Y consists of the following steps: 1) **Transcript extraction and segmentation:** We normalize our dataset so that we convert

[5] https://escriba.aosfatos.org/en/

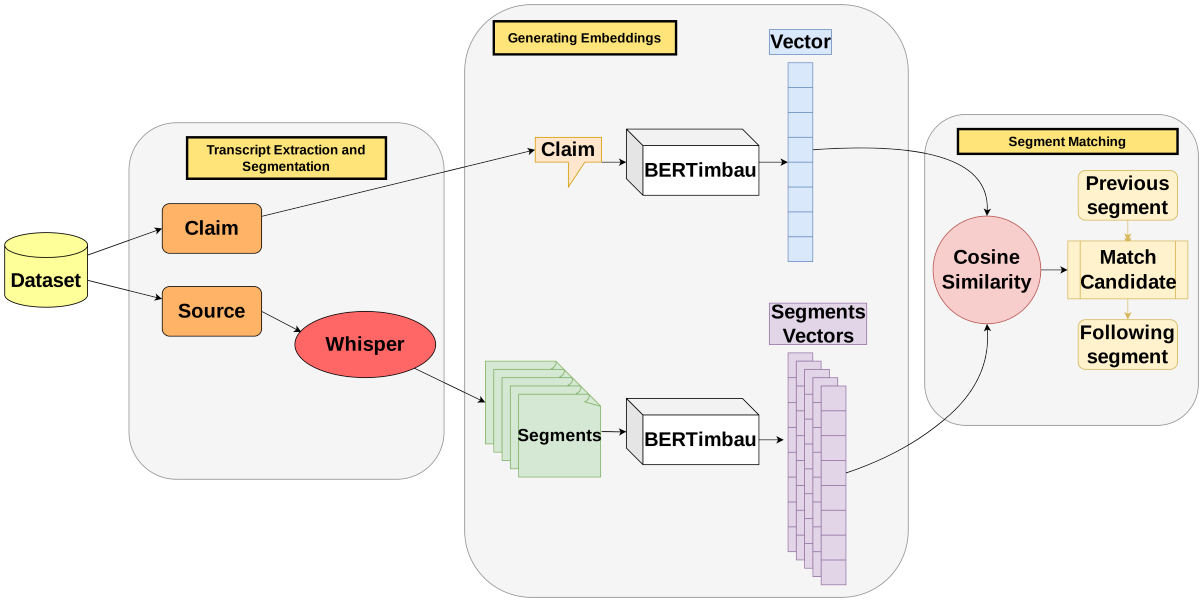


Figure 3.1: Overview of our methodology regarding the BOL4Y dataset

each video to a textual transcript, which we segment into pieces; 2) **Segment embeddings generation:** We convert the segmented textual data from the transcript into dense embeddings using a BERT-based model; 3) **Perform segment matching:** We semantically match the segments with the annotated false claims from AosFatos using the segment embeddings; and 4) **Classification:** We perform a segment-level classification to identify segments sharing false information, essentially solving the misinformation spans detection task. We present an overview of our methodology in Figure 3.1. Below, we elaborate on these steps and our experimental setup.

3.2.1.1 Transcript Extraction & Segmentation

Our approach to misinformation span detection in videos leverages the transcriptions of videos' audios. To extract transcriptions from videos, we use OpenAI's Whisper [80], a state-of-the-art speech recognition model, on the audio of each video in our dataset. Whisper takes as input an audio file and generates a textual transcription. Although Whisper cannot provide word-level timestamps [68], it can segment audio into transcribed segments (i.e., parts of the transcription) of at most 30-second windows. We applied Whisper to the 525 video files in our dataset and extracted their textual transcripts. Note that the transcripts provided by Escriba are already segmented by AosFatos' Escriba service.

3.2.1.2 Generating Segment Embeddings

Having converted our dataset into textual information (i.e., textual transcripts) divided into segments, our next step is to align the transcribed segments with manually annotated fact-checks provided by aosfatos. To this end, we use a state-of-the-art transformer-based model trained and tailored for Brazilian Portuguese. Specifically, we use BERTimbau [95], a BERT-based model [27] that is pre-trained on the Brazilian Web as Corpus (BrWac) [105], a large Brazilian Portuguese corpus. The model was downloaded from the HuggingFace repository, [6] and we use the base model that yields embeddings of 768 dimensions. Moreover, we use the SentenceTransformers [82] implementation to retrieve the embeddings from the mean pooling of the language model. In a nutshell, BERTimbau takes as input the textual information included in a transcript segment and generates a dense vector representation (*embedding*); these embeddings are the foundation for matching segments that share misinformation as they allow us to assess the similarity of transcript segments and fact-checked claims.

3.2.1.3 Performing Segment Matching

Here, we aim to identify the transcript segments that contain misinformation claims, as fact-checked by professional journalists. To leverage these fact-checks as positive (i.e., misinformation) labels in our dataset, we compare all transcript segments from a given video to the actual fact-check available for that video. This is an integral part of our methodology as it allows us to create an annotated dataset of segments that share misinformation and segments that do not. To achieve this, we perform the following procedure: We use BERTimbau to extract embeddings for each segment of each video transcript in our dataset (see Section 3.2.1.2). We also compute embeddings for each false claim (see *Claim* field in Section 3.2). Recall that each false claim is associated with one video in the dataset. Then, we compare the claim's embedding to all video segments' embeddings using cosine similarity. For each false claim, we consider the segment with the highest cosine similarity as the top candidate to be examined. This part allows for the identification of segments that potentially share misinformation, as they share textual similarities with the known false claims. Given that a false claim may span into multiple transcription segments, we also extract the segments before and after the segment with the highest cosine similarity for further examination.

[6] https://huggingface.co/neuralmind/bert-base-portuguese-cased

After identifying segments that potentially share false claims, we perform a manual annotation process to verify that they are indeed sharing false claims. We use the open-source text annotation tool Doccano[7,8] to speed up the annotation procedure. For the annotation procedure, we focus on pairs of top candidate segments and false claims with a cosine similarity of 0.7 or higher. We selected this threshold after manual examinations that showed that pairs with a cosine similarity of 0.7 or below were not semantically similar. After applying this threshold and selecting all pairs of top candidate segments/false claims with cosine similarity higher than 0.7, we end up with 2,996 pairs that we annotate. For each claim, we prompt the annotator to flag which of the three selected segments comprise the claim. We choose a comprehensive approach and flag every segment that has at least one word that is part of the claim. We also add flags to i) signal if more segments are needed to capture the whole claim, ii) signal if there is a mistranscription (i.e., some words of one or more segments seem to be mistranscribed); iii) None of the segments shown match the fact-checked claim.

We perform additional rounds of segment matching with the instances flagged as missing a part of the claim, adding more segments before and after the already flagged segments. Two annotators matched 2,373 claims from the initial 6,685 claims listed by AosFatos, with the two annotators disagreeing only on 18 cases, which were discussed and removed from the dataset, resulting in 2,355 total segments with a 99.24% agreement rate between annotators. Afterward, for each matched claim, we concatenate the segments composing that claim into one and consider that concatenation as a positive example in further steps. The reason for merging these segments is to ensure that the full context of the claim is considered. In cases where a claim is spread across multiple segments, each segment on its own might not contain enough information to determine if it is misinformative. Note that these claims come from a subset of the initial set of downloaded videos: 430 videos out of the initial 525 and 108 out of the initial 121 transcriptions from Escriba, totaling 538 unique sources.

Finally, since our goal is to model our problem as a segment classification task, we need segments that do not share false claims (i.e., negative examples). To do this, we treat all segments that are not matched or annotated as negative examples (i.e., segments that do not share false claims). Using this approach, we end up with 336,855 segments that we treat as negative examples.

[7] https://github.com/doccano/doccano

[8] An example of Doccano's interface is available in Appendix B

3.3 EI22 dataset

We also release a second expert-annotated dataset gathered by AosFatos. AosFatos fact-checked a set of videos posted on YouTube and privately shared them with us. The dataset comprises 78 fact-checked videos of electoral fraud claims made by voters during the 2022 Brazilian presidential election. We refer to this dataset as **E**lection **I**ntegrity **22**, shortened to EI22. In total, EI22 has 77 videos and 1997 segments, of which 78 are misinformative claims. The 77 videos are of varying lengths, come from voters' own recordings, and are entirely separate from the videos on the BOL4Y dataset.

3.3.1 Building EI22

AosFatos provided us with a list of videos that comprise EI22, which contained the links to the videos and timestamps of the misinformative claims. We again employed Whisper, which transcribed the audio into segments. We then selected the segments that comprised the timestamps of the claims, relying on the expertly annotated timestamps.

3.4 Classification

To investigate the feasibility of automatically detecting false segments in video transcripts, we employ two models pre-trained with Brazilian Portuguese as bases for our classifiers: BERTimbau and PTT5. We use BERTimbau, which we already use for extracting segment embeddings, and PTT5 [13], which was also pre-trained on the BrWac collection and is based on the T5 architecture [85]. For each of these models, we use a classification head with a softmax activation that provides us with a probability of the segment sharing false claims or not for each segment. Note that for the classification, we elected to use PTT5 in addition to BERTimbau to compare how the selection of the underlying Transformer architecture (i.e., encoder-only vs. encoder-decoder) affects the classification performance.

3.5 Experimental Setup

Here, we provide more details on our experimental setup, including information about the dataset preparation, training and evaluation, and temporal and cross-dataset experiments.

3.5.1 Dataset Preparation.

Our BOL4Y dataset is highly imbalanced: 2,355 positive instances (i.e., segments sharing false claims) and 336,885 negative instances (i.e., segments that do not share false claims). This substantial class imbalance impacts classification performance. Hence, we evaluate classification performance using various configurations by randomly undersampling[9] the negative examples in the training dataset. In particular, we use the following ratios: 1-to-1 (i.e., balanced training set across classes), 1-to-10, 1-to-25, 1-to-50, 1-to-75, 1-to-100, and the full dataset (2.3K positive and 336K negative examples). It is relevant to mention that undersampling is applied only to the training set, with both validation and test sets being kept intact.

3.5.2 Dataset Variations.

AosFatos published the list of claims on their website. However, they may present editing by their journalists to correct grammatical errors or, in some cases, to add some context within brackets. There are also additional challenges in working with transcriptions, such as noisy audio, poor transcription, and imperfect speech. Considering the issues mentioned, the edited version of the claim might differ from what we find in the transcripts.

Therefore, we have created an alternative version of the dataset in which we have replaced the false claims found in the transcripts with the version released by the jour-

[9]In addition to random undersampling, other techniques are available to undersample a dataset. We choose the random alternative due to being a simpler and easily reproducible alternative. Other options are available in the imbalanced-learn package: https://imbalanced-learn.org/stable/under_sampling.html

nalist. Hence, we will refer to the variation of the dataset with the claims written by the journalist as the "Edited" version and the version with claims extracted from transcriptions as the "Original" version. For reference, we provide an example of what a claim looks like in the original and edited datasets in Table 3.1: note that the version from the edited dataset has context added in brackets. Given the polished nature of the edited dataset, we aim to provide insights into the challenges of working with transcriptions for misinformation span detection and how ill-formated claims might be detrimental to performance. We aim to assess how the quality of the transcripts affects the classification performance when considering the misinformation span detection task.

Table 3.1: Example of claim in the original and edited datasets

Dataset Variation	**Bolsonaro's Claim**
Original	"He built three hydroelectric power plants abroad"
Edited	"He [Lula, Brazil's former president] built three hydroelectric power plants abroad"

3.5.3 Training and Evaluation.

We use the HuggingFace implementations of the BERTimbau[10] and PTT5[11] models, which we fine-tune for our dataset variations using a Nvidia T4 GPU. The HuggingFace implementations contain a classification head that produces the output prediction from the generated embeddings of the model. We perform classification with 5-fold cross-validation. For each fold, we divide the dataset into five equal portions; three are used for training, one for validation, and one for testing. The validation set is used in an early-stopping approach, as we use the model from the epoch that best performed in the validation set. We train the models for three epochs and use default parameters from their implementation. Then, we assess the performance of classifiers and the impact of training set sizes on evaluation results. The undersampled variations also give us an insight into how classifiers can be implemented and used in the wild, as a bigger dataset also implicates using more resources to train models. We follow the above procedure considering different undersampling ratios over the original and edited datasets.

[10] https://huggingface.co/neuralmind/bert-base-portuguese-cased
[11] https://huggingface.co/unicamp-dl/ptt5-base-portuguese-vocab

3.5.3.1 Metrics

We evaluate our experiments using Precision and Recall for class 1 (misinformation), Balanced Accuracy, and Macro-f1. We argue that these metrics comprehensively overview our models' performance across settings. Overall, we consider Macro-F1 our main metric due to class imbalance.

We also argue that false negatives are more relevant than false positives in our setup. The main goal of our methodology is to aid fact-checkers. Considering this, false positives, that is, legitimate claims predicted as false, will be double-checked by journalists with no added harm. However, false negatives will go unnoticed, as they will not be flagged as false correctly, resulting in a potentially more problematic and harmful outcome.

3.5.4 Sliding Window Experiments.

We also conduct experiments in a temporal manner to evaluate the real-world feasibility of detecting misinformation in future data. Specifically, we investigate whether models trained on past months' data can accurately predict subsequent months' misinformation. We perform two settings: 1) fixed training and 2) expanding training windows.

In the first setting, the training and test sets span fixed periods (6 months for training and one month for testing). We progressively move the testing window forward by one month. In the second experiment, the test window remains fixed for one month, but the training window expands with each iteration. During training, we use the most recent month as validation data for both settings. We train each variation for three epochs and select the model with the best performance on the validation set. The temporal experiments aim to investigate: If we train models on data from a given period in months, can we accurately predict misinformation in future months?

3.5.5 Cross-dataset performance.

We also perform a cross-dataset test, training models with BOL4Y, and testing on EI22, striving to assess cross-domain performance. Recall that these datasets pertain to different contexts: BOL4Y relates to false claims made by Bolsonaro while EI22 relates

to electoral fraud claims made by voters. We argue that the context difference between the datasets allows us to perform cross-domain experiments. We train models for three epochs with BERTimbau and PTT5 using multiple undersampling ratios.

Apart from releasing additional data (i.e., the EI22 dataset) for the task, we aim to provide insights into data representativeness of data and misinformative claims; we also want to assess how models trained in one dataset perform when tested in another dataset of claims made by different speakers, further discussed in Section 3.6.1, effectively showcasing the feasibility of the task in a real-world scenario.

3.6 Results

We now go over the results of our proposed experiments

3.6.1 Classification Performance

3.6.1.1 Original Dataset

Table 3.2 shows the results for our classifiers regarding all considered variations of our training dataset. Recall that, due to the considerable amount of data, we leverage undersampling variations of our dataset as our training set while maintaining the same test sets for all experiments. We consider six positive-to-negative example ratios (1-to-1, 1-to-10, 1-to-25, 1-to-50, 1-to-75, and 1-to-100) along the full dataset when undersampling our training set. We implement a 5-fold cross-validation approach and report average values on a video level, i.e., we compute metrics for every video in every fold and report average value for five folds. We compare results with the Macro F1 score due to class imbalance, along with class-balanced accuracy, and precision and recall for Class 1 (misinformation).

The BERTimbau classifier trained on the full version of our dataset is outperformed by all undersampled versions. The same happens for the PTT5 classifier trained on the full dataset. These results motivate us to exclude the full version of the dataset from further experiments due to its poor performance and high training time. The BERTimbau-based classifiers match or outperform the PTT5 ones when comparing the same training sets

Table 3.2: Classification results for our dataset

	BERTimbau (Full)	BERTimbau (1-to-1)	BERTimbau (1-to-10)	BERTimbau (1-to-25)	BERTimbau (1-to-50)	BERTimbau (1-to-75)	BERTimbau (1-to-100)
Balanced Accuracy	0.55	0.82	0.78	0.75	0.68	0.69	0.62
Macro F1	0.56	0.49	0.63	0.67	0.66	0.68	0.63
Precision (Class 1)	0.21	0.09	0.24	0.35	0.38	0.43	0.35
Recall (Class 1)	1.00	0.75	0.94	0.97	0.99	0.99	1.00
	PTT5 (Full)	**PTT5 (1-to-1)**	**PTT5 (1-to-10)**	**PTT5 (1-to-25)**	**PTT5 (1-to-50)**	**PTT5 (1-to-75)**	**PTT5 (1-to-100)**
Balanced Accuracy	0.54	0.81	0.76	0.70	0.64	0.60	0.58
Macro F1	0.54	0.49	0.61	0.64	0.62	0.60	0.58
Precision (Class 1)	0.15	0.08	0.20	0.30	0.29	0.28	0.27
Recall (Class 1)	1.00	0.76	0.94	0.97	0.99	0.99	1.00

regarding Macro F1. The BERTimbau-based classifier trained on the smallest training set (1-to-1 ratio) yields the best-balanced accuracy value, achieving a 0.82 score, although with poorer recall, precision, and Macro F1. Regarding Macro F1, BERT (1-to-75) yields the best performance overall, with a Macro F1 score of 0.68. We see a positive impact on performance when varying the undersampling ratio, with better results than training models with the full dataset. This shows that training models in a full dataset setting can be counterproductive in addition to being more costly. Overall, these results highlight that misinformation span detection is challenging, with modern classifiers based on state-of-the-art language models achieving an F1 score of up to 0.68.

3.6.1.2 Edited Dataset

We also propose an analysis of classification using an alternative version of our dataset where we consider the claims as edited by the journalist. To provide context, we initially performed a sentence-matching task to locate fact-checked claims within video transcriptions. To better understand the challenges of using transcriptions as input for classification, we have created an alternative version of the dataset. In this version, we replaced the transcribed claims (which served as positive examples) with the original claims as presented by AosFatos' journalists. These original claims are more refined and polished in comparison.

Comparatively, we see that unpolished claims (i.e., the original dataset) degrade performance, which might be attributed to the noisy nature of transcriptions as they, for example, can replicate speech imperfections from the original audio. We see an increase in performance when using the edited version of the dataset (See Table 3.3) when comparing models trained in datasets with different undersampling ratios, which showcases the difficulty of working with transcriptions. Particularly for PTT5, the best-performing version is now the 1-to-75 undersampled version instead of 1-to-25, as shown in Table 3.2. Overall, we find that the Edited dataset shows how ill-formatted claims can be detrimental to

performance.

Table 3.3: Classification results for the Edited version of our dataset.

	BERTimbau (1-to-1)	BERTimbau (1-to-10)	BERTimbau (1-to-25)	BERTimbau (1-to-50)	BERTimbau (1-to-75)	BERTimbau (1-to-100)
Balanced Accuracy	0.91	0.92	0.88	0.85	0.85	0.81
Macro F1	0.60	0.73	0.78	0.81	0.81	0.81
Precision (Class 1)	0.21	0.39	0.52	0.62	0.65	0.68
Recall (Class 1)	0.87	0.97	0.98	0.99	0.99	1.00
	PTT5 (1-to-1)	**PTT5 (1-to-10)**	**PTT5 (1-to-25)**	**PTT5 (1-to-50)**	**PTT5 (1-to-75)**	**PTT5 (1-to-100)**
Balanced Accuracy	0.90	0.90	0.88	0.81	0.80	0.77
Macro F1	0.58	0.71	0.75	0.76	0.79	0.76
Precision (Class 1)	0.19	0.37	0.46	0.54	0.63	0.60
Recall (Class 1)	0.85	0.97	0.98	0.99	1.00	0.99

3.6.2 Temporal Analyses

We also conduct experiments where we partition the dataset by organizing the claims according to the specific months when Bolsonaro made them. We aim to gain insights into the practicality of deploying misinformation detection models in real-world scenarios where future data is inaccessible. This experiment will help us evaluate the robustness of our models in predicting and detecting future misinformation, focusing on the task of misinformation span detection.

We base our temporal analysis on the best-performing models in Table 3.2 regarding Macro F1 scores, namely BERT-75 and T5-25. Then, we propose two separate temporal studies for each: 1) a fixed training span of six months, hereafter referenced as **Walk-Forward** and 2) an increasing training span, hereafter referenced as **Expand**, starting with six months.

Figure 3.2 shows the distribution of Bolsonaro's false claims over time and important milestones of his presidency. Notably, false claims increased during the COVID-19 pandemic, starting to lower after the first quarter of 2022 and swiftly growing nearer to the presidential elections, when Bolsonaro faced his biggest political opponent, Brazil's then-former president, Lula.

We train all models for three epochs and, considering the last month of the training set as a validation set, choose the best version using early stopping. Note that due to the lack of claims in June 2019 (see Figure 3.2), we could not use it as a validation or test set, yielding null scores for June 2019 (test) and July 2019 (validation). In both settings ("Walk Forward" and "Expand"), we test models on the month chronologically after the month of the validation set. We consider the unedited dataset and report Macro F1 scores monthly in Figure 3.3. Results show values ranging from 0.5 to 0.8, with the

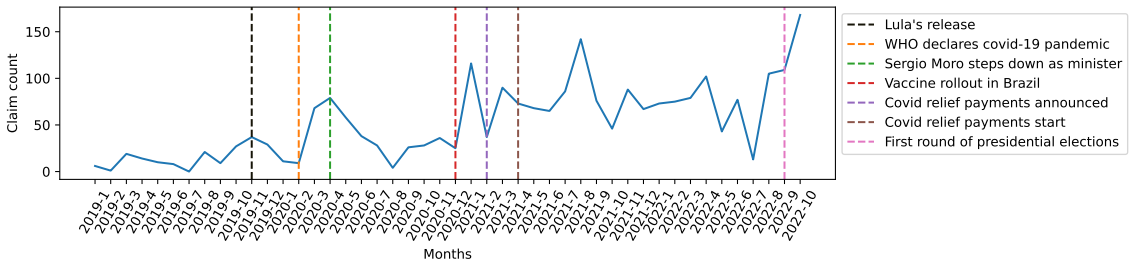
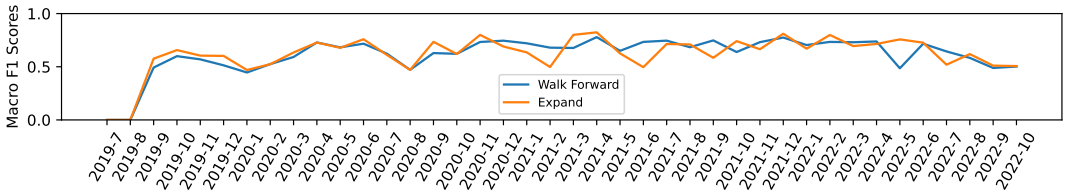
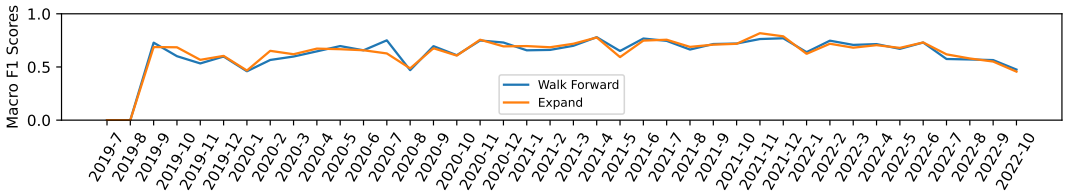


Figure 3.2: Monthly sum of misinformation claims. Vertical lines signal important events during Bolsonaro's administration.

overall highest score on April 2021 (BERTimbau - Expand), the month after the start of Covid relief payments. For all settings, we observe a decrease in Macro F1 scores during the second semester of 2022, even for the "Expand" approaches, which are trained on all previous months. Overall, we also note that PTT5 yields more consistent performance across settings, generating similar results for "Walk Forward" and "Expand" in contrast to BERTimbau.



(a) BERTimbau 75 - 6 Month Training Period



(b) PTT5 25 - 6 Month Training Period

Figure 3.3: Temporal analysis of the performance of our classifiers.

3.6.2.1 Cross-dataset performance

Table 3.4 shows the cross dataset experiment results. We trained models with BOL4Y and tested on EI22. We varied the undersampling ratio, achieving the best result (Macro F1 score of 0.72) with the 1-to-10 ratio for both models. Our results point to

cross-dataset effectiveness, which is crucial in dealing with misinformation in a realistic setting.

Table 3.4: Macro F1 scores for cross-dataset performance

	1-to-1	1-to-10	1-to-25	1-to-50	1-to-75	1-to-100
BERTimbau	0.64	0.71	0.62	0.62	0.58	0.61
PTT5	0.64	0.71	0.63	0.57	0.59	0.56

3.6.3 Factors Affecting Performance (BOL4Y)

Here, we conduct additional analyses to understand how the classification performance is affected by various factors over the BOL4Y dataset, including the source of the transcription, the quality of the transcription, and the topic of the false claim. We choose to perform these analyses on the BOL4Y as it is a much larger dataset with claims from multiple topics, in contrast to EI22.

3.6.3.1 Noise Scores

As mentioned previously, there are multiple sources of confusing factors when dealing with transcriptions: noisy audio, poor transcription, and imperfect speech. First, we wanted to quantify the impact of noise on classification performance, so we calculated the Spectral Flatness score [32] for all of our videos' audio. Spectral flatness (or tonality coefficient) measures how much a sound resembles white noise, as opposed to a pure tone. A high spectral flatness (equal to 1.0) indicates that the sound has a flat spectrum, similar to white noise. The score distribution is shown in Figure 3.4, highlighting that most videos have clear audio, except for a few outliers.

Nevertheless, we assess correlation between metrics and spectral flatness. Table 3.5 shows the Pearson correlation between spectral flatness and Macro F1 scores for the BERT-75 variation. We find no relevant correlation between Macro F1 and Spectral Flatness values, possibly due to the nature of the videos: many come from interviews and live streams, mostly recorded in quiet environments.

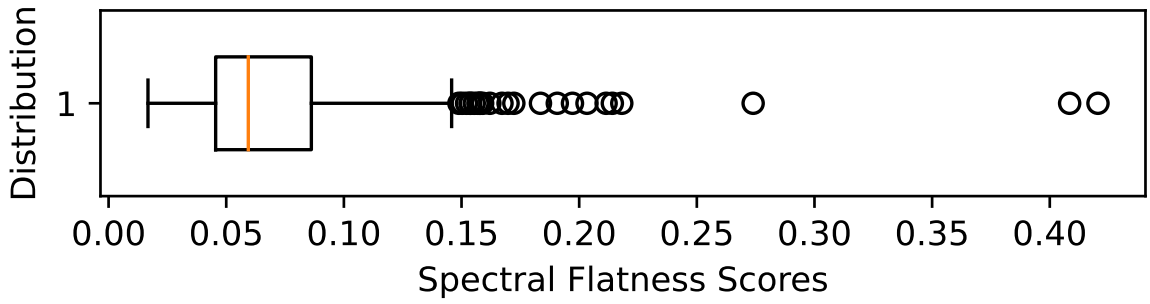


Figure 3.4: Spectral Flatness Scores.

Table 3.5: Correlation between performance and noise.

	Spectral Flatness Correlation
Balanced Accuracy	-0.045
Macro F1	-0.046
Precision (Class 1)	-0.054
Recall (Class 1)	-0.01

3.6.3.2 Transcription Source

Recall that the transcriptions in our dataset come from 1) videos we downloaded and transcribed using Whisper and 2) transcriptions provided directly by AosFatos using their automated, proprietary transcription tool, "Escriba". To assess the possible impacts of transcription sources, we present the distribution of scores divided by sources: Figure 3.5 shows distributions of Macro F1 scores for both. There's a clear difference between Whisper and Escriba transcriptions, validated through a statistical test of means (Mann-Whitney U, $p < 0.0001$), which motivates us to understand possible causes with an additional analysis regarding editing distance between original (i.e., as written by AosFatos' journalists) and transcribed claims.

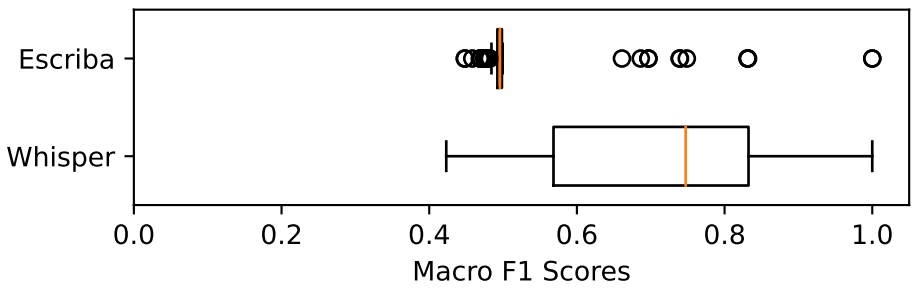


Figure 3.5: Macro F1 score distribution by transcription source.

3.6.3.3 Editing Distance

Here, we investigate how the classification performance changes based on the difference between the segment containing a claim and the fact-checked claim from AosFatos. To do this, first, we calculate the editing distance between the original claim (i.e., as written by AosFatos' journalists) and the claim within the transcription. Then, for each video, we compute the average editing distance between all claims in said video. Finally, we calculate the correlation between the Macro F1 score and the average editing distance for all sources, finding a weak Pearson correlation (-0.37) between the two variables. This result points to an impact of the properties of transcribed text in classification performance. We hypothesize that this may be due to added context provided by journalists in some claims through information in brackets, as exemplified in Section 3.5.2.

3.6.3.4 Themes

Finally, we provide insights into theme-wise performance: a single video often contains multiple claims, and these claims can cover a range of themes that AosFatos' journalists annotate. To start, we compute the frequency of these themes within our dataset and order them from most to least frequent. Figure 3.6 illustrates how the Macro F1 scores are distributed among the top 7 most prevalent themes, each occurring at least 100 times in our dataset. We notice that the performance is generally consistent across different themes, except for claims related to Congress and the Judicial System, which exhibit poorer performance. For each theme, we add the number of times they occur in parenthesis on the x-axis of Figure 3.6, and although "Judicial System" and "Congress" have distributions skewed to lower Macro F1 values, they have similar frequencies to "Environment" and "Elections". These results point to the possible effects of different themes in misinformation span detection in videos, and further analyses are left for future work.

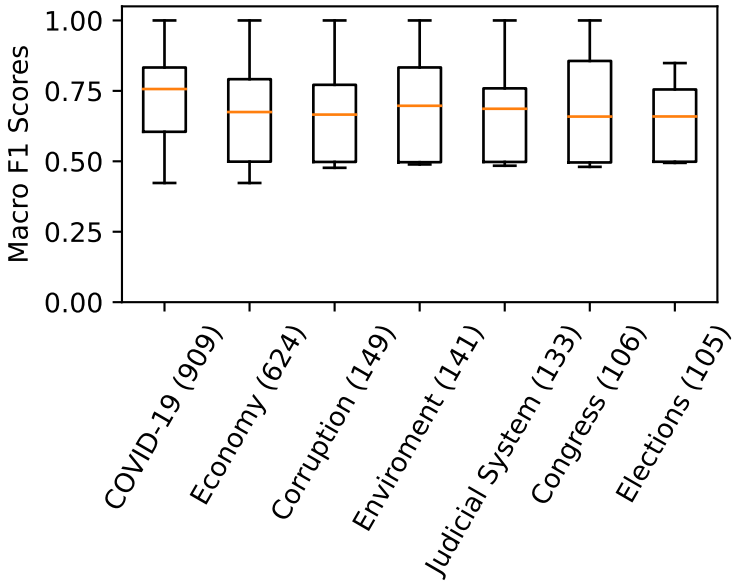


Figure 3.6: Macro F1 score distribution for the seven most common themes in our dataset. In parenthesis, the number of times that each theme occurs.

3.7 Employing LLMs

3.7.1 Fine-tuning

Our work on misinformation span detection, so far, showed the feasibility of the task, which can be improved in future work, particularly using larger language models, such as LLaMa [99, 100] or ChatGPT [1]. Specifically, LLMs leverage architectural advancements that yield high-quality, contextually relevant responses without extensive fine-tuning, contrasting with smaller models such as BERT or T5.

Initially, we tested a fine-tuned classifier built with the 13B[12] variation of LLaMa 2 and trained for three epochs, selecting the best-performing model using a validation set. We used a LoRA [45] approach for fine-tuning for better efficiency, which only adjusts a small percentage of weights (6% in our case). We set out to test classifiers with varying undersampling ratios, as we did for BERTimbau and PTT5-based classifiers. We tested a LLaMa-based classifier trained using an undersampled version of BOL4Y with 1-to-1, 1-to-10, and 1-to-100 ratios, with results shown in Table 3.6. For comparison purposes, we also reiterate 1) the results for the best-performing versions of BERTimbau (1-to-75) and PTT5 (1-to-25) and 2) respective models trained with a 1-to-1 ratio version of BOL4Y. However, our LLaMa-based classifiers cannot outperform the much less costly

[12] 13 billion parameters

BERT and T5-based approaches. In summary, the tradeoff between computational cost and performance was not worthwhile in our experiments, so we turned to alternatives that leverage LLMs in potentially more efficient ways via in-context learning.

Table 3.6: Results for LLaMa classifier. We also recall the best models for BERTimbau and PTT5 for comparison purposes

	BERTimbau (1-to-75)	**PTT5 (1-to-25)**	**BERTimbau (1-to-1)**	**PTT5 (1-to-1)**	**LLaMa 2 (1-to-1)**	**LLaMa 2 (1-to-10)**	**LLaMa 2 (1-to-100)**
Balanced Accuracy	0.69	0.70	0.82	0.81	0.77	0.75	0.49
Macro F1	0.68	0.64	0.49	0.49	0.45	0.63	0.32
Precision (Class 1)	0.43	0.30	0.09	0.08	0.08	0.24	0.02
Recall (Class 1)	0.99	0.97	0.75	0.76	0.69	0.96	0.61

3.7.2 In-context Learning

Another way to leverage LLMs' extended capabilities is through in-context learning (ICL) [11], that is, learning from a few examples in the context of the prompt. Dong et al. [31] distinguish between supervised learning and in-context learning: "Different from supervised learning requiring a training stage that uses backward gradients to update model parameters, ICL does not conduct parameter updates and directly performs predictions on the pretrained language models. The model is expected to learn the pattern hidden in the demonstration and accordingly make the right prediction.". Figure 3.7 illustrates an example of misinformation classification via in-context learning.

In this section, we show a proof-of-concept classification experiment using in-context learning. The goal is to leverage the capabilities of large language models, namely LLaMa 2, to perform detection through ICL.

Although ICL is more efficient than fine-tuning an LLM, it is still a costly approach. So, for this experiment, we use a subset of the BOL4Y: We consider all 2,355 false claims and select the same amount of non-misinformative claims, totaling 4710 segments for classification.

As discussed in Chapter 2, the idea of using in-context learning is to remove the costs of fine-tuning LLMs by providing demonstrations as part of the prompt, which can be effective for many tasks, such as sentiment analysis. In our task, we provide segments of videos as demonstrations followed by their label and then prompt the model with a new segment for its label. We performed multiple rounds of prompt engineering and settled on the prompt shown in Figure 3.8.

We also select the demonstrations used in our prompts, as shown in Figure 3.9. Recall that each video has at least one theme, with many having two, as expertly an-

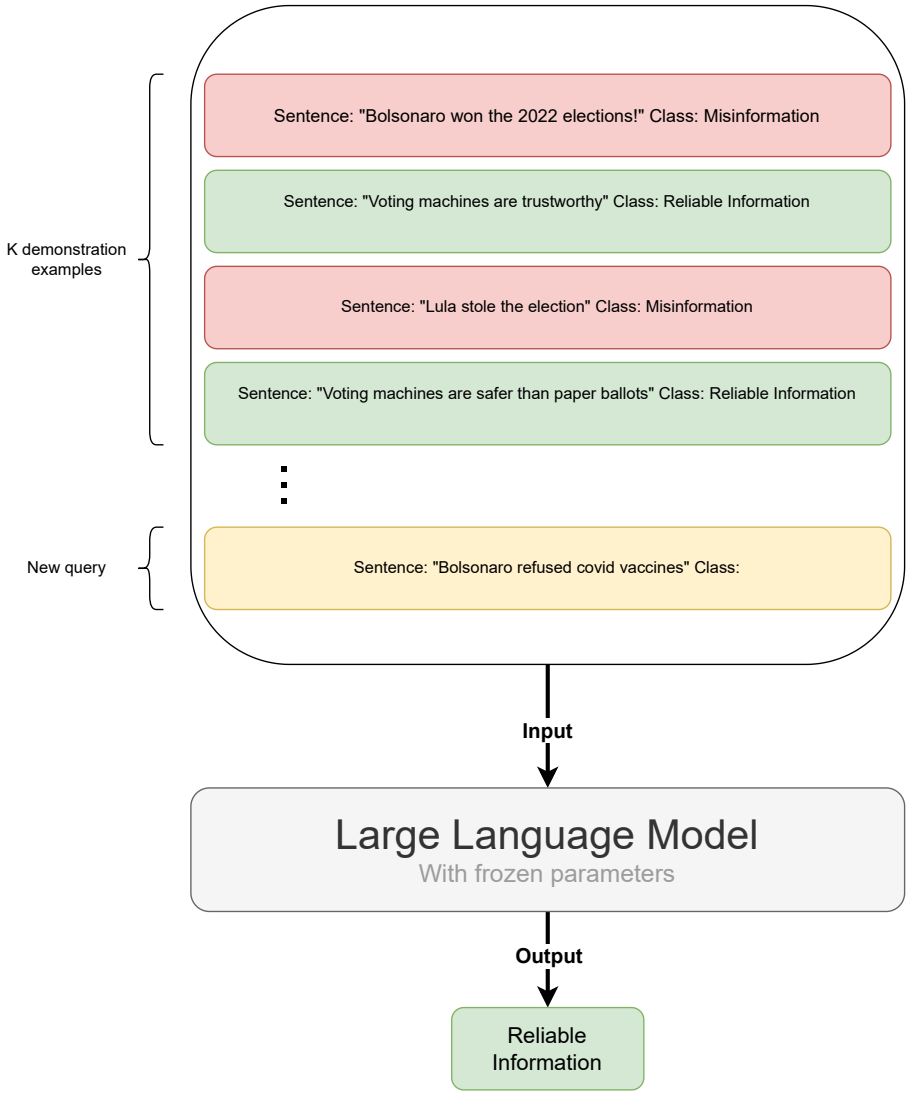


Figure 3.7: Example of misinformation detection via in-context learning

notated by AosFatos' journalists. Then, given a video segment we want to classify, we retrieve the theme of the video that contains said segment. Afterward, we randomly select other videos of the same theme and extract both positive and negative examples as demonstrations. If a video has more than one theme, we select half of the demonstrations from videos of each theme. We then feed this prompt to the LLaMa model.

We experimented with two different LLaMa 2 variations, 13B and 70B, on an A100 GPU with 80GBs. We utilized the 70B variation with 8-bit quantization and the 13B version with full precision. Our goal was to perform a comprehensive assessment of our setup as we tested two model variations, one of which is one of the largest open-weight models available. Moreover, we also wanted to compare performance across model variations.

Recall that we consider two classes, misinformation and non-misinformation, which

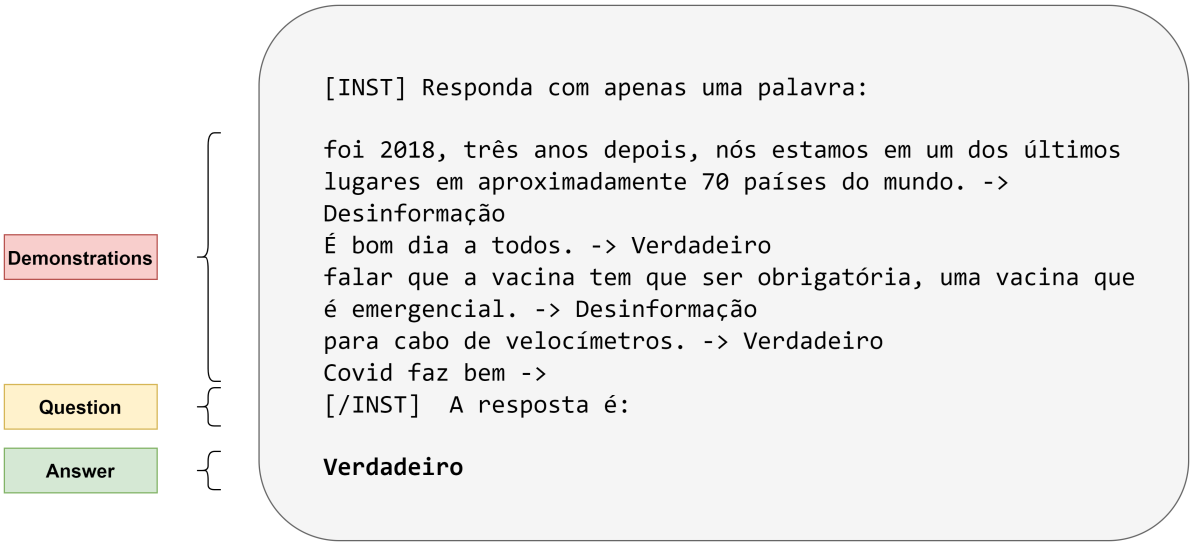


Figure 3.8: Example of one prompt used.

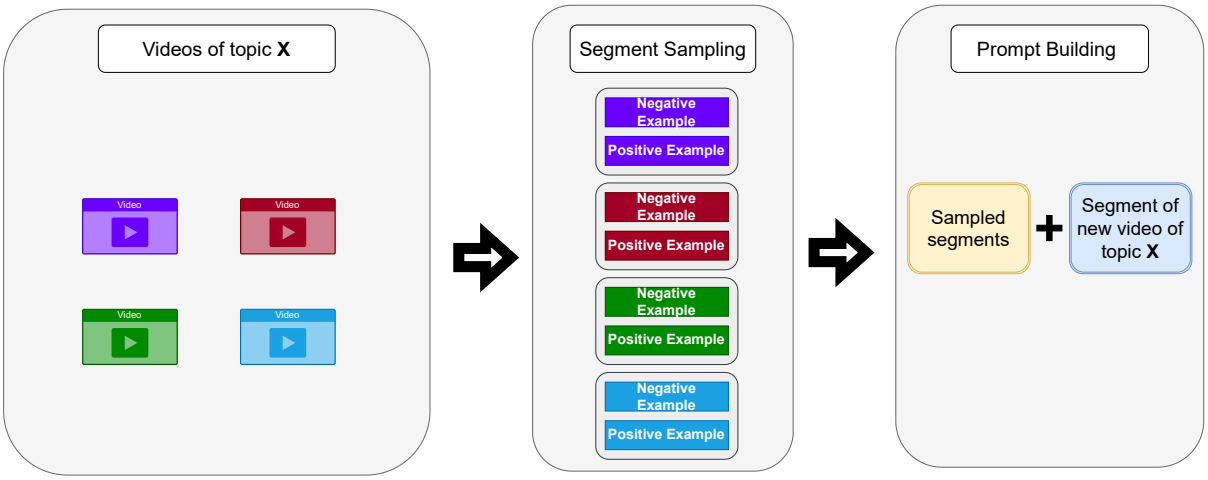


Figure 3.9: Building our prompts. We select segments, both negative

we use when building the demonstrations in our prompts, as exemplified in Figure 3.9. However, as LLMs are generative models, sometimes they do not adhere to using only one of the two words and output unrelated words or gibberish. We aimed to quantify our results using the same metrics used in Section 3.6.1. So, whenever our model outputs unrelated words, we treat it like an error: we check the correct label for that instance and attribute the opposite class as our prediction. For example, if our model outputs gibberish, but the proper label is "misinformation", we consider our prediction "non-misinformation". We do this to provide a clear evaluation setup that allows us to use classic machine learning metrics.

We lowercase all outputs and remove all whitespaces and punctuation to normalize the output tokens, focusing on standardizing the outputs, which will be useful when grouping equivalent predictions further on. Table 3.7 shows the number of predictions split by class across the two LLaMa variations. 70B has higher prompt adherence than

	13B	**70B**
Misinformation	3239	3333
Non-misinformation	475	1203
Other	996	174

Table 3.7: Predictions grouped by class for each model

13B; that is, most answers follow the command given, outputting one of the two correct classes, an expected result given the difference in model size. Table 3.8 displays each model's top 40 most frequent outputs. We highlight that, besides more output variation, the 13B model also shows a few terms in English instead of Portuguese, which does not occur in the 70B variation, which we hypothesize is also due to different capabilities due to model size.

A few terms that can be read as equivalents to our two classes are also present in both models. We highlight "falso" (line 4) and "verdade" (line 5) for 13B and "verdade" (line 17) and "verdadeiro" (line 18) for 70B. We choose not to compute performance metrics considering these instances as they require a manual evaluation, which defeats the purpose of our automated misinformation detection approach.

	LLaMa 13B		LLaMa 70B	
	Output	**# of occurrences**	**Output**	**# of occurrences**
1	desinformação	3239	desinformação	3333
2	verdadeiro	475	verdadeiro	1203
3	não	307	não	27
4	falso	112	okay	6
5	verdade	78	bolsonaro	5
6	desinformado	76	impossível	4
7	sim	38	agenda	3
8	based on the statements	33	exagero	3
9	incorrect	21	brasil	3
10	falsidade	16	falso	2
11	based on the information	13	caro	2
12	engano	12	respeito	2
13	ironia	8	impeachment	2
14	okay	7	ok	2
15	desinformados	7	desinformado	2
16	falsa	6	ironia	2
17	desastre	6	verdade	2
18	based on the text	6	verdadeira	2
19	sure here	5	paraguai	1
20	nenhum	5	para	1
21	zero	5	percepção	1
22	brasil	5	perguntar	1
23	falsário	4	polémico	1
24	nenhuma	4	parabéns	1
25	okay here are	4	política	1
26	sure here are	4	participação	1
27	bolsonaro	3	a	1
28	liberdade	3	orientação	1
29	lula	3	oportunidade	1
30	transparência	3	opinião	1
31	heres the	3	okay i	1
32	desinformadas	3	obrigado	1
33	desconhecido	3	não não	1
34	com base nas inform	3	norte	1
35	errado	2	normalidade	1
36	opinião	2	negócios	1
37	engana	2	negou	1
38	com base nas res	2	okay aqui	1
39	false	2	projeto	1
40	here are the answers	2	políticos	1

Table 3.8: Top 40 most frequent outputs segmented by model variation. Whitespace and punctuations have been removed.

We then evaluate classification performance, shown in Table 3.9, using precision and recall for class 1 (misinformation), balanced accuracy, and macro-f1. As expected, the 70B model performs better than the 13B version, with both higher precision and recall

for class 1, resulting in a higher macro-F1. However, we observed poor performance that did not compare with the best BERT or T5-based models shown in Table 3.2.

	13B	70B
Balanced Accuracy	0.37	0.51
Macro F1	0.31	0.48
Precision (Class 1)	0.42	0.51
Recall (Class 1)	0.08	0.27

Table 3.9: Results for proof-of-concept experiment with ICL

3.8 Limitations

Next, we discuss some limitations of our methodology on misinformation span detection. First, although we use Whisper, a high-quality transcription model, audio transcriptions can still be noisy data, and transcription models depend heavily on the audio quality to yield good results. Additionally, Whisper does the segmentation process automatically and on a sentence level. Currently, word-level segmentation is not supported in Whisper [68]. Some transcriptions come from Escriba, AosFatos' proprietary transcription service that does not disclose details on implementation.

Additionally, although hard annotation was done by professional fact-checkers (journalists), and the task in our study was very straightforward (check the similarity of two segments), segment matching has a subjective component which can be a limitation.

Finally, our data is focused only on the Brazilian context, which is restricted to the Portuguese language. Representativeness is an important but challenging issue in any empirical study such as ours. We argue that the Brazilian context is relevant to the field of misinformation, and our data covers a wide range of themes highly exploited by misinformation campaigns over four years [98, 83]. For instance, our data includes Bolsonaro's livestreams, which are organized periodically and used to construct narratives along different topics that would favor the former Brazilian president.

3.9 Discussion

The work presented in this chapter aimed to define the task of misinformation span detection, showcase initial efforts in solving the task, and evidence additional factors that might affect performance.

We proposed two novel datasets for the task (i.e., BOL4Y and EI22) and aimed to assess multiple classification setups. The great imbalance between classes (i.e., misinformation and non-misinformation) in the BOL4Y dataset prompted us to test undersampled versions of it, that is, undersampling the majority class (i.e., non-misinformation) in the multiple classification setups. We employed two language models fine-tuned for Brazilian Portuguese in our tests, BERTimbau and PTT5, over multiple undersampling ratios. We found that different models have distinct undersampling ratios that work best, pointing to the optimal parameters being model-specific, with our best model reaching a 0.68 Macro F1 score.

We also performed tests with the Edited dataset to understand if added context and less noisy data could foster more effective classification. We found this to be the case, as unpolished claims degrade performance, with much better performance in the edited dataset for both models. Furthermore, we proposed temporal experiments that aimed to assess the robustness of these models over time in two distinct setups, finding that both models yield fair results, which can be useful in a real-world setting. We also performed a cross-dataset experiment by training models on BOL4Y and testing on the EI22 dataset, achieving a Macro F1 of 0.71, pointing to cross-dataset robustness. Moreover, we aimed to identify additional factors affecting performance, discovering variations based on the transcription method, the claim's theme, and the level of text editing, which can guide future efforts in misinformation span detection.

Finally, we performed proof-of-concept tests with LLaMa 2, classifying a subset of the BOL4Y dataset via 1) fine-tuning and 2) in-context learning. We found that the fine-tuning approach is not justifiable, as it does not point to a significant increase in performance and demands much higher computational costs than training a BERT or T5-based classifier. Additionally, our ICL strategy, albeit much less costly than its fine-tuning counterpart, does not yield good results, with low macro-F1 and accuracy. We experimented with two LLaMa variations, including the largest available when writing this thesis, and found the performance of both models to be subpar. We hypothesize that this might be due to these pre-trained models' intrinsic limitations, in addition to the nature of our data and task.

3.10 Future Work

In future work, we want to incorporate more features in the classification pipeline, such as metrics related to abusive language, such as toxicity, profanity, and inflammatory scores; this can be achieved with tools such as Perspective API,[13] which we believe can provide valuable data for improved classification approaches. We are also convinced that explicability frameworks, such as SHAP [52] or LIME [84], can help us probe into our classifiers, evidencing their inner workings and exploring error cases.

Additionally, we also want to experiment with additional features extracted from the video itself, such as facial expressions, as we believe that these can add valuable information to our classifiers. To achieve this, we can leverage vision LLMs, such as LLaMa 3.2.[14] Finally, since LLMs usually have a data cutoff of several months before their release, additional rounds of fine-tuning can help add new knowledge to these models, potentially improving performance.

[13] https://perspectiveapi.com/

[14] https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

Chapter 4

Conclusion

Online misinformation is a rampant, multi-faceted issue that affects online environments, taking on various forms, such as memes, images, and content disseminated through social networks, dedicated communities, and messaging apps like WhatsApp and Telegram. It spreads through various mediums, encompassing audio, text, images, and video content. Given the relevance and ubiquity of the problem, we aimed to propose mitigation solutions for misinformative content.

In this work, we presented the first effort to explore the problem of misinformation span detection in videos, focusing on videos flagged as misinformative by expert fact-checkers. We presented the first effort to explore the problem of misinformation span detection in videos. In addition to determining whether a video contains misinformation, we also identify the specific part (span) of the video where it occurs. We investigated multiple setups to assess the challenges related to effective misinformation span detection. We achieved promising results, with our best model yielding an F1 score of 0.68, indicating the feasibility of the task. Furthermore, we built the first two datasets for misinformation span detection and made them available to the scientific community as one of the contributions of our work; our datasets provide completely novel data for a new, unexplored task. We also assessed cross-dataset performance, achieving an F1 score of 0.71 when training with the BOL4Y dataset and testing it on the EI22 dataset with both BERTimbau and PTT5; this points to effective detection despite misinformative claims coming from different sources. Furthermore, we perform proof-of-concept experiments with LLaMa 2, a large language model, with both fine-tuning and ICL.

Finally, others can replicate the pipeline proposed in this paper to build new datasets for different contexts, further improving automatic misinformation detection. We hope our methodology for misinformation span detection can be used to develop other applications to assist fact-checkers and reduce the time spent on misinformation detection in videos by pinpointing potential fact-checking points. Also, we argue that identifying the spans of misinformation within videos can assist social media operators in providing additional context to viewers when a false claim occurs. For instance, they can include warning labels with additional context regarding a false claim as an overlay on a video when a false claim is made.

Finally, our work comes at a critical time for digital platforms. Initiatives like the Digital Services Act (DSA) regulation have emerged as significant steps forward in regulating digital spaces, aiming to ensure safer and more responsible online environments through effective content moderation. Such initiatives highlight the need for more robust automatic moderation tools, and we hope our work can improve these efforts.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Seyi Akiwowo, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem. Proceedings of the fourth workshop on online abuse and harms. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 2020.

[3] Omar Ali, Nancy Scheidt, Alexander Gegov, Ella Haig, Mo Adda, and Benjamin Aziz. Automated detection of racial microaggressions using machine learning. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pages 2477–2484. IEEE, 2020.

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

[5] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion proceedings of the 2019 world wide web conference*, pages 1100–1105, 2019.

[6] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. *First monday*, 21(11-7), 2016.

[7] Mary Blankenship. How misinformation spreads through twitter. 2020.

[8] Lia Bozarth and Ceren Budak. Market forces: Quantifying the role of top credible ad servers in the fake news ecosystem. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 83–94, 2021.

[9] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1664–1674, 2019.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell,

Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[12] Kevin Matthe Caramancion. Understanding the association of personal outlook in free speech regulation and the risk of being mis/disinformed. In *2021 IEEE World AI IoT Congress (AIIoT)*, pages 0092–0097, 2021. doi: 10.1109/AIIoT52608.2021.9454212.

[13] Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*, 2020.

[14] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In Tsung-Hsien Wen, Asli Celikyilmaz, Zhou Yu, Alexandros Papangelis, Mihail Eric, Anuj Kumar, Iñigo Casanueva, and Rushin Shah, editors, *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlp4convai-1.5. URL https://aclanthology.org/2020.nlp4convai-1.5.

[15] Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. I feel offended, don't be abusive! implicit/explicit messages in offen-

sive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, 2020.

[16] Caroline CAUFFMAN and Catalina GOANTA. A new order: The digital services act and consumer protection. *European Journal of Risk Regulation*, 12(4):758–774, 2021. doi: 10.1017/err.2021.8.

[17] Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. "subverting the jewtocracy": Online antisemitism detection using multimodal deep learning. In *Proceedings of the 13th ACM Web Science Conference 2021*, pages 148–157, 2021.

[18] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

[19] Matteo Cinelli, Andraž Pelicon, Igor Mozetič, Walter Quattrociocchi, Petra Kralj Novak, and Fabiana Zollo. Dynamics of online hate and misinformation. *Scientific Reports*, 11(1):22083, Nov 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-01487-w. URL https://doi.org/10.1038/s41598-021-01487-w.

[20] Isobelle Clarke and Jack Grieve. Dimensions of abusive language on twitter. In *Proceedings of the first workshop on abusive language online*, pages 1–10, 2017.

[21] Victor Claussen. Fighting hate speech and fake news. the network enforcement act (netzdg) in germany in the context of european legislation. *Rivista di diritto dei media*, 3:1–27, 2018.

[22] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska De Jong. Improving cyberbullying detection with user context. In *Advances in Information*

Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings 35, pages 693–696. Springer, 2013.

[23] Alina Darmstadt, Mick Prinz, and Oliver Saal. The murder of keira: Misinformation and hate speech as far-right online strategies. 2019.

[24] Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. Hatemm: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1014–1023, 2023.

[25] Sourya Dipta Das, Ayan Basak, and Saikat Dutta. A heuristic-driven uncertainty based ensemble framework for fake news detection in tweets and news articles. *Neurocomputing*, 491:607–620, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.12.037. URL https://www.sciencedirect.com/science/article/pii/S0925231221018750.

[26] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL https://aclanthology.org/2020.acl-main.372.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. ACL'19*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[29] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.

[30] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30, 2015.

[31] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[32] Shlomo Dubnov. Generalization of spectral flatness measure for non-gaussian linear processes. *IEEE Signal Processing Letters*, 11(8):698–701, 2004.

[33] Azza El-Masri, Martin J Riedl, and Samuel Woolley. Audio misinformation on whatsapp: A case study from lebanon. *Harvard Kennedy School (HKS) Misinformation Review*, 2022.

[34] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

[35] Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. Understanding and countering stereotypes: A computational approach to the stereotype content model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 600–616, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.50. URL https://aclanthology.org/2021.acl-long.50.

[36] Tamara Fuchs and Fabian Schäfer. Normalizing misogyny: hate speech and verbal abuse of female politicians on japanese twitter. In *Japan forum*, volume 33, pages 553–579. Taylor & Francis, 2021.

[37] Anastasia Giachanou and Paolo Rosso. The battle against online harmful information: The cases of fake news and hate speech. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3503–3504, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3412169. URL https://doi.org/10.1145/3340531.3412169.

[38] Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso. Multimodal multi-image fake news detection. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 647–654. IEEE, 2020.

[39] Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233, 2017.

[40] Matthew Hannah. Qanon and the information dark age. *First Monday*, 2021.

[41] Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. Arcov19-rumors: Arabic covid-19 twitter dataset for misinformation detection. *arXiv preprint arXiv:2010.08768*, 2020.

[42] K. Hayawi, S. Shahriar, M.A. Serhani, I. Taleb, and S.S. Mathew. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public Health*, 203:23–30, 2022. ISSN 0033-3506. doi: https://doi.org/10.1016/j.puhe.2021.11.022. URL https://www.sciencedirect.com/science/article/pii/S0033350621004534.

[43] Sarah Hewitt, Thanassis Tiropanis, and Christian Bokhove. The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, pages 333–335, 2016.

[44] Rui Hou, Verónica Pérez-Rosas, Stacy Loeb, and Rada Mihalcea. Towards automatic detection of misinformation in online medical videos. In *2019 International conference on multimodal interaction*, pages 235–243, 2019.

[45] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[46] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[47] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. Measuring misinformation in video search platforms: An audit study on youtube. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), may 2020. doi: 10.1145/3392854. URL https://doi.org/10.1145/3392854.

[48] Raj Jagtap, Abhinav Kumar, Rahul Goel, Shakshi Sharma, Rajesh Sharma, and Clint P George. Misinformation detection on youtube using video captions. *arXiv preprint arXiv:2107.00941*, 2021.

[49] Mladen Karan and Jan Šnajder. Cross-domain detection of abusive language online. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 132–137, 2018.

[50] Alireza Karduni, Isaac Cho, Ryan Wesslen, Sashank Santhanam, Svitlana Volkova, Dustin L Arendt, Samira Shaikh, and Wenwen Dou. Vulnerable to misinformation? verifi! In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 312–323, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362726. doi: 10.1145/3301275.3302320. URL https://doi.org/10.1145/3301275.3302320.

[51] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, WWW '19, page 2915–2921, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313552. URL https://doi.org/10.1145/3308558.3313552.

[52] Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. Bert meets shapley: Extending shap explanations to transformer-based classifiers. In *Proceedings of the EACL hackashop on news media content analysis and automated report generation*, pages 16–21, 2021.

[53] Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 1621–1622, 2013.

[54] Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1131. URL https://aclanthology.org/D19-1131.

[55] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

[56] Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. Benchmarking natural language understanding services for building conversational agents. In *Increasing naturalness and flexibility in spoken dialogue interaction: 10th international workshop on spoken dialogue systems*, pages 165–183. Springer, 2021.

[57] Zhiwei Liu, Kailai Yang, Qianqian Xie, Christine de Kock, Sophia Ananiadou, and Eduard Hovy. Raemollm: Retrieval augmented llms for cross-domain misinformation detection using in-context learning based on emotional information. *arXiv preprint arXiv:2406.11093*, 2024.

[58] Caio Machado, Beatriz Kira, Vidya Narayanan, Bence Kollanyi, and Philip Howard. A study of misinformation in whatsapp groups with a focus on the brazilian presidential elections. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1013–1019, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366755. doi: 10.1145/3308560.3316738. URL https://doi.org/10.1145/3308560.3316738.

[59] Golshan Madraki, Isabella Grasso, Jacqueline M. Otala, Yu Liu, and Jeanna Matthews. Characterizing and comparing covid-19 misinformation across languages, countries and platforms. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 213–223, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383134. doi: 10.1145/3442442.3452304. URL https://doi.org/10.1145/3442442.3452304.

[60] Ahmed Redha Mahlous and Ali Al-Laith. Fake news detection in arabic tweets during the covid-19 pandemic. *International Journal of Advanced Computer Science and Applications*, 12(6):778–788, 2021.

[61] Alexandre Maros, Jussara Almeida, Fabrício Benevenuto, and Marisa Vasconcelos. Analyzing the use of audio messages in whatsapp groups. In *Proceedings of The Web Conference 2020*, WWW '20, page 3005–3011, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380070. URL https://doi.org/10.1145/3366423.3380070.

[62] Alexandre Maros, Jussara M. Almeida, and Marisa Vasconcelos. A study of misinformation in audio messages shared in whatsapp groups. In Jonathan Bright, Anastasia Giachanou, Viktoria Spaiser, Francesca Spezzano, Anna George, and Alexandra Pavliuc, editors, *Disinformation in Open Online Media*, pages 85–100, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87031-7.

[63] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 173–182, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362023. doi: 10.1145/3292522.3326034. URL https://doi.org/10.1145/3292522.3326034.

[64] Breno Matos, Rennan C. Lima, Jussara M. Almeida, Marcos André Gonçalves, and Rodrygo L. T. Santos. On the presence of abusive language in mis/disinformation. In Frank Hopfgartner, Kokil Jaidka, Philipp Mayr, Joemon Jose, and Jan Breitsohl, editors, *Social Informatics*, pages 292–304, Cham, 2022. Springer International Publishing. ISBN 978-3-031-19097-1.

[65] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

[66] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416.

[67] Xiaoli Nan, Yuan Wang, and Kathryn Thier. Health misinformation. 2021.

[68] Openai. Getting time offsets of beginning and end of each word · openai/whisper · discussion #3. URL https://github.com/openai/whisper/discussions/3#discussioncomment-3703465. Accessed on 12/10/2023.

[69] Demetris Paschalides, Alexandros Kornilakis, Chrysovalantis Christodoulou, Rafael Andreou, George Pallis, Marios Dikaiakos, and Evangelos Markatos. Check-it: A plugin for detecting and reducing the spread of fake news and misinformation on the web. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, page 298–302, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369343. doi: 10.1145/3350546.3352534. URL https://doi.org/10.1145/3350546.3352534.

[70] Umaru A Pate and Adamkolo Mohammed Ibrahim. Fake news, hate speech and nigeria's struggle for democratic consolidation: A conceptual review. *Handbook of research on politics in the computer age*, pages 89–112, 2020.

[71] Parth Patwa, Mohit Bhardwaj, Vineeth Guptha, Gitanjali Kumari, Shivam Sharma, Srinivas PYKL, Amitava Das, Asif Ekbal, Md Shad Akhtar, and Tanmoy Chakraborty. Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In Tanmoy Chakraborty, Kai Shu, H. Russell Bernard, Huan Liu, and Md Shad Akhtar, editors, *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 42–53, Cham, 2021. Springer International Publishing. ISBN 978-3-030-73696-5.

[72] John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. Semeval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pages 59–69, 2021.

[73] Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. The surprising performance of simple baselines for misinformation detection. In *Proceedings of the Web*

Conference 2021, WWW '21, page 3432–3441, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450111. URL https://doi.org/10.1145/3442381.3450111.

[74] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio Arechar, Dean Eckles, and David Rand. Understanding and reducing the spread of misinformation online. *ACR North American Advances*, 2020.

[75] Konstantin Pogorelov, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, and Johannes Langguth. Wico text: A labeled dataset of conspiracy theory and 5g-corona misinformation tweets. In *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks*, OASIS '21, page 21–25, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386326. doi: 10.1145/3472720.3483617. URL https://doi.org/10.1145/3472720.3483617.

[76] Amit Praseed, Jelwin Rodrigues, and P. Santhi Thilagam. Hindi fake news detection using transformer ensembles. *Engineering Applications of Artificial Intelligence*, 119: 105731, 2023. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2022.105731. URL https://www.sciencedirect.com/science/article/pii/S0952197622007217.

[77] Jiameng Pu, Neal Mangaokar, Lauren Kelly, Parantapa Bhattacharya, Kavya Sundaram, Mobin Javed, Bolun Wang, and Bimal Viswanath. Deepfake videos in the wild: Analysis and detection. In *Proceedings of the Web Conference 2021*, pages 981–992, 2021.

[78] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE international conference on data mining (ICDM)*, pages 518–527. IEEE, 2019.

[79] Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14444–14452, 2023.

[80] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

[81] Shaina Raza and Chen Ding. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4):335–362, May 2022. ISSN 2364-4168. doi: 10.1007/s41060-021-00302-z. URL https://doi.org/10.1007/s41060-021-00302-z.

[82] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proc. EMNLP '19*. ACL, 11 2019. URL https://arxiv.org/abs/1908.10084.

[83] Julio C. S. Reis, Philipe Melo, Kiran Garimella, Jussara M. Almeida, Dean Eckles, and Fabrício Benevenuto. A Dataset of Fact-Checked Images Shared on WhatsApp During the Brazilian and India Elections. *ICWSM*, 2020.

[84] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[85] Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J Liu, Sharan Narang, Wei Li, and Yanqi Zhou. Exploring the limits of transfer learning with a unified text-to-text transformer. 2019.

[86] Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576, 2020.

[87] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3(1):80–87, 2019.

[88] Punyajoy Saha, Binny Mathew, Kiran Garimella, and Animesh Mukherjee. "short is the road that leads from fear to hate": Fear speech in indian whatsapp groups. In *Proceedings of the Web Conference 2021*, WWW '21, page 1110–1121, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450137. URL https://doi.org/10.1145/3442381.3450137.

[89] Gaurav Sahu, Pau Rodriguez, Issam H Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. Data augmentation for intent classification with off-the-shelf large language models. *arXiv preprint arXiv:2204.01959*, 2022.

[90] Akrati Saxena, Wynne Hsu, Mong Li Lee, Hai Leong Chieu, Lynette Ng, and Loo Nin Teow. Mitigating misinformation in online social network with top-k debunkers and evolving user opinions. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 363–370, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370240. doi: 10.1145/3366424.3383297. URL https://doi.org/10.1145/3366424.3383297.

[91] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE International*

Conference on Big Data (Big Data), pages 899–908, 2021. doi: 10.1109/BigData52589.2021.9671928.

[92] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9(1):4787, Nov 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-06930-7. URL https://doi.org/10.1038/s41467-018-06930-7.

[93] Bhuvanesh Singh and Dilip Kumar Sharma. Predicting image credibility in fake news over social media using multi-modal approach. *Neural Computing and Applications*, 34(24):21503–21517, 2022.

[94] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 39–47, 2019. doi: 10.1109/BigMM.2019.00-44.

[95] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.

[96] Wen-Ying Sylvia Chou, Anna Gaysynsky, and Joseph N Cappella. Where we go from here: health misinformation on social media, 2020.

[97] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. Defining "fake news" a typology of scholarly definitions. *Digital journalism*, 6(2):137–153, 2018.

[98] Cristina Tardaguila, Fabricio Benevenuto, and Pablo Ortellado. Fake news is poisoning brazilian politics. whatsapp can stop it, 2018. URL https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html.

[99] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[100] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann,

Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[101] Ciprian-Octavian Truică and Elena-Simona Apostol. Misrobærta: Transformers versus misinformation. *Mathematics*, 10(4), 2022. ISSN 2227-7390. doi: 10.3390/math10040569. URL https://www.mdpi.com/2227-7390/10/4/569.

[102] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[103] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. Polarization and fake news: Early warning of potential misinformation targets. *ACM Trans. Web*, 13(2), mar 2019. ISSN 1559-1131. doi: 10.1145/3316809. URL https://doi.org/10.1145/3316809.

[104] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[105] Jorge Wagner, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. The brwac corpus: A new open resource for brazilian portuguese. 05 2018.

[106] Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policymaking, 2017.

[107] Shao Yi Liaw, Fan Huang, Fabricio Benevenuto, Haewoon Kwak, and Jisun An. Younicon: Youtube's community of conspiracy videos. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):1102–1111, Jun. 2023. doi: 10.1609/icwsm.v17i1.22218. URL https://ojs.aaai.org/index.php/ICWSM/article/view/22218.

[108] Savvas Zannettou, Mai Elsherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. Measuring and characterizing hate speech on news&websites. In *12th*

ACM Conference on Web Science, WebSci '20, page 125–134, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379892. doi: 10.1145/3394231.3397902. URL https://doi.org/10.1145/3394231.3397902.

[109] Alexandros Zervopoulos, Aikaterini Georgia Alvanou, Konstantinos Bezas, Asterios Papamichail, Manolis Maragoudakis, and Katia Kermanidis. Deep learning for fake news detection on twitter regarding the 2019 hong kong protests. *Neural Computing and Applications*, 34(2):969–982, 2022.

[110] Baybars Örsek. Ifcn is heartened by a nobel peace prize nomination for the work of the global fact-checking community, Oct 2021. URL https://www.poynter.org/fact-checking/2021/ifcn-is-heartened-by-a-nobel-peace-prize-nomination-for-the-work-of-the-global-fact-checking-community/.

Appendix A

Statistics on Videos and Comments From the BOL4Y Dataset

Considering the BOL4Y dataset, we initially downloaded 525 videos, from which 460 have comments from users available. Recall that we could only match claims from 430 videos, as described in Section 3.2.1.3. However, all videos we downloaded were flagged as containing misinformation, so we chose to show statistics related to the 460 videos with user comments, which all come from YouTube. In total, we gathered 1,738,946 comments.

A.1 Videos

Figure A.1 shows the distribution of videos regarding upload date across the four years of Bolsonaro's presidency, with 2021 being the year with the most videos.

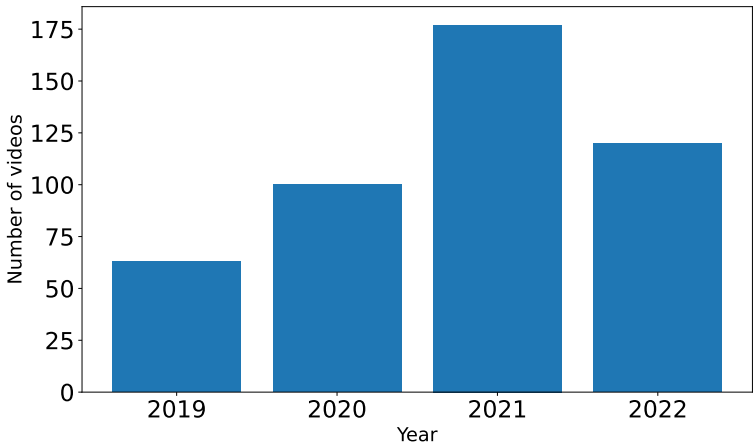


Figure A.1: Distribution of videos over the years.

Then, Table A.1 displays channels that posted the videos, namely the top 10 with the most videos uploaded. We highlight that the channel that uploaded the most videos was Bolsonaro's official YouTube channel, with other two Bolsonaro-affiliated channels in

the top 10: Bolsonaro TV, and Carlos Bolsonaro, with the latter being one of Bolsonaro's son's channel. Additionally, from the 460 videos, 174 were live streams, which was one of the ways Bolsonaro used to communicate with his voter base.

Table A.1: Top 10 channels with the most videos in respect to the total 460

Channel	**Video Count**
Jair Bolsonaro	196
Foco do Brasil	108
Bolsonaro TV	29
CanalGov	15
Poder360	12
Carlos Bolsonaro	6
SBT News	6
CNN Brasil	6
Band Jornalismo	5
Os Pingos nos Is	5

Moreover, Table A.2 displays the categories YouTube attributed to each video. Expectedly, the vast majority of videos pertain to the News & Politics category.

Table A.2: Video count by category

Video Category	**Count**
News & Politics	409
Entertainment	28
People&Blogs	4
Travel&Events	2
Education	7
Science&Technology	1
Gaming	6
Film&Animation	2
Sports	1

We then move to general statistics regarding the videos. As evidenced by Figure A.2, the views distribution is heavy-tailed, with most videos having less than 2.5M views. A similar pattern appears in Figure A.3, with a heavy-tailed distribution and most videos receiving under 250,000 likes. Overall, we see that the videos we gathered have high engagment, with many views and likes.

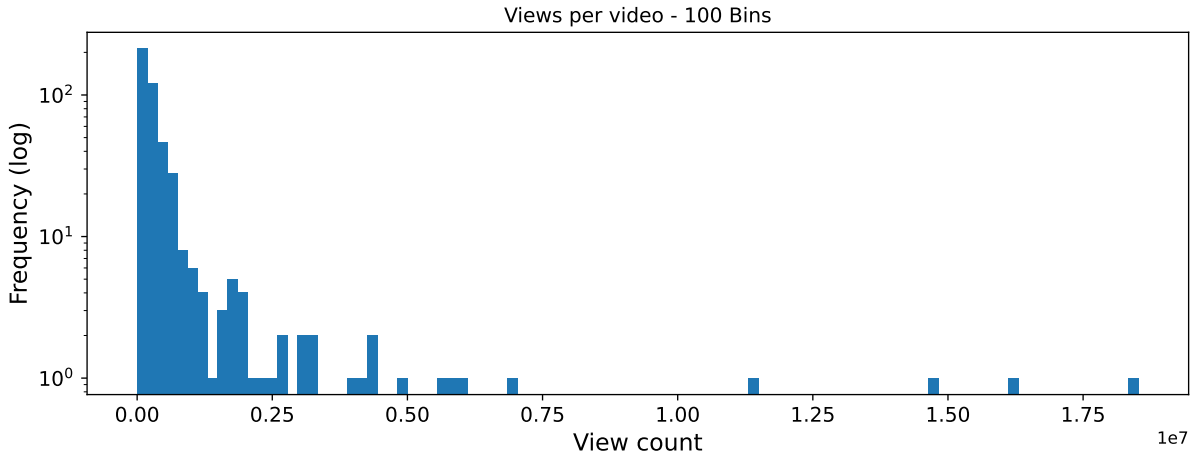


Figure A.2: Views per video. X-axis values should be multiplied by 1e7

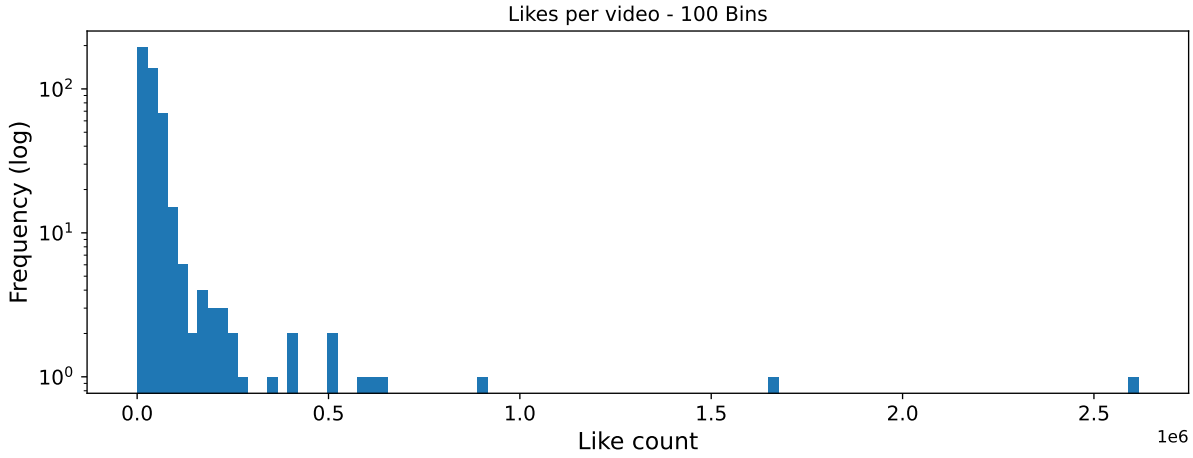


Figure A.3: Likes per video. X-axis values should be multiplied by 1e6

Furthermore, we provide insight into the length of videos (in seconds) in Figure A.4. Most videos lie before the 5000 second mark (approximately 83 minutes), evidencing that these videos are usually long in duration

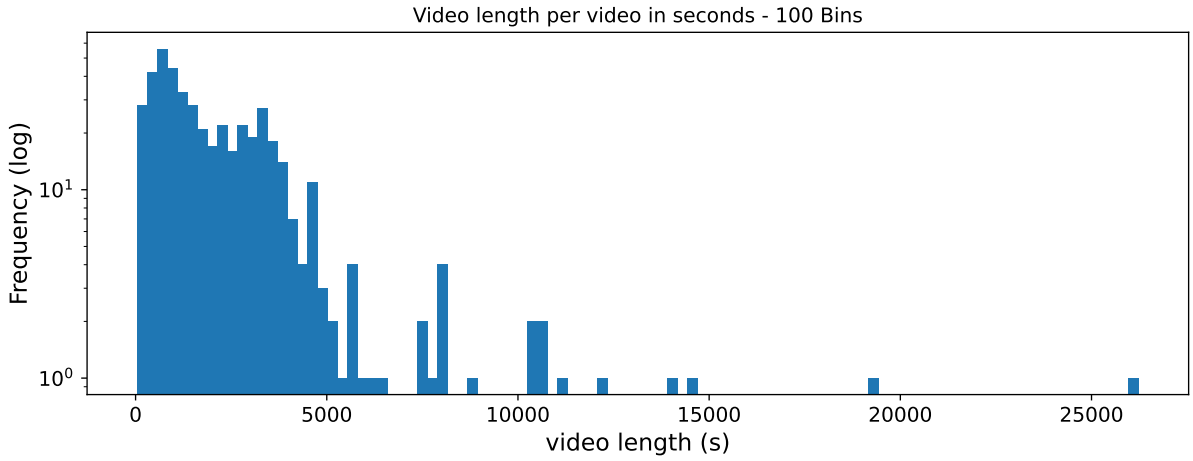


Figure A.4: Video length

Finally, Figure A.5 shows the distribution of comments per video, reinforcing the high engagement they have.

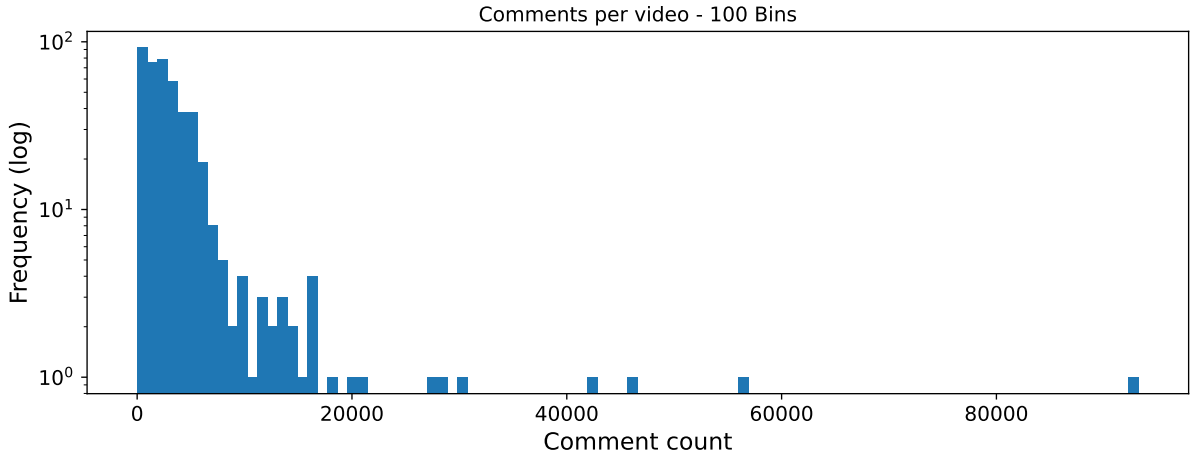


Figure A.5: Distribution of comments per video. 100 bins. Y-axis in log scale

We now turn to taking a deeper look into the comments' content in Section A.2

A.2 Comments

Figure A.6 shows the distribution of comments in regard to their word count, with most ranging from 1 to 1000 words.

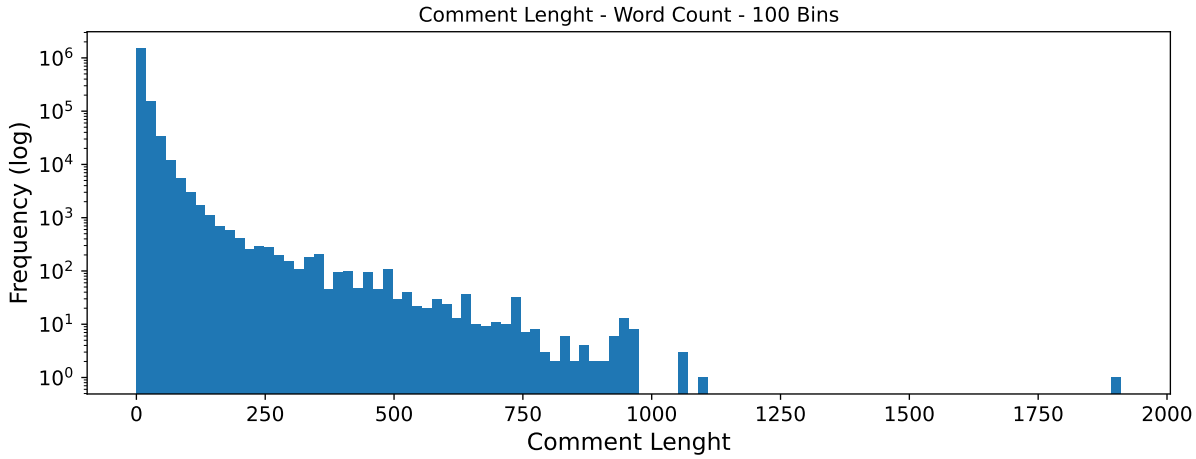


Figure A.6: Comment word count

Furthermore, we once again employed Perspective API to the 1,738,946 comments and display the distribution for six attributes: Toxicity, Severe Toxicity, Identitiy Attack, Insult, Profanity, and Threat. We show results in Figure A.7 and highlight that 25% of comments present scores equal to or higher than 0.6 for toxicity, identity attack, and insult, pointing to significant hostility in the comments.

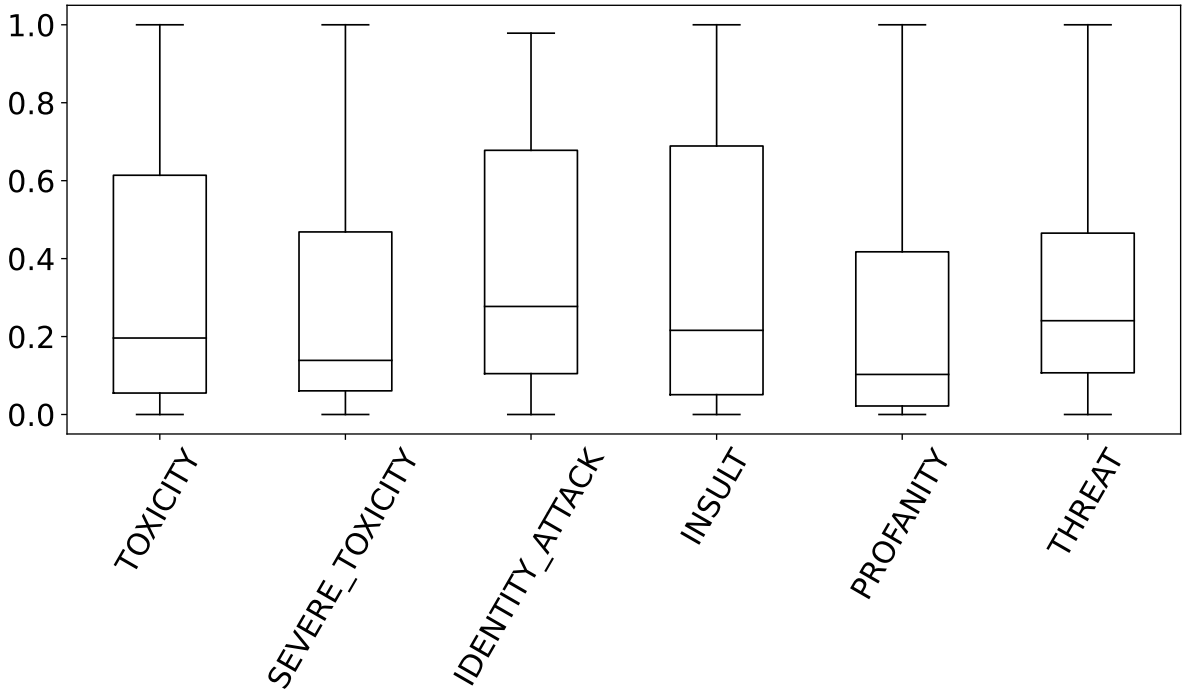


Figure A.7: Distribution of Perspective Attributes

Finally, we analyze the top 20 most used emojis, displayed in Figure A.8.[1]. The most used emoji used is the Brazilian flag, followed by emojis of endorsement. We also highlight emojis that might be used by anti-Bolsonaro commenters, such as the bull/cow related emojis, often used to antagonize Bolsonaro's voters.

Emoji	Occurence
🇧🇷	633337
👏	311776
🙏	167337
💚	80062
😂	70569
💛	65013
🤣	62939
👍	61987
🐂	50611
❤	45461
🙌	40732
😎	32083
🐃	28244
🐄	28057
🐂	25795
🤮	21186
💙	20639
💪	20519
😍	18450
🎉	15661

Figure A.8: Top 20 most used emojis

[1]We choose to display the emojis table as a figure due to LaTeX's poor emoji support

Appendix B

Doccano

Figure B.1 shows an example of Doccano's interface during the segment matching task. We present the claim and the candidate segments related to it, and the user must flag which segments (if any) comprise said claim.

Figure B.1: Example of Doccano's interface