UMA ANÁLISE DE MENSAGENS DE ÁUDIO COMPARTILHADAS EM GRUPOS DO WHATSAPP

ALEXANDRE MAROS

UMA ANÁLISE DE MENSAGENS DE ÁUDIO COMPARTILHADAS EM GRUPOS DO WHATSAPP

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Jussara Marques de Almeida
Coorientador: Marisa Affonso Vasconcelos

Belo Horizonte
Novembro de 2020

ALEXANDRE MAROS

AN ANALYSIS OF AUDIO MESSAGES SHARED IN WHATSAPP GROUPS

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

Advisor: Jussara Marques de Almeida
Co-Advisor: Marisa Affonso Vasconcelos

Belo Horizonte
November 2020

Maros, Alexandre.

M354u Uma análise de mensagens de áudio compartilhadas em grupos do whatsapp [manuscrito]. / Alexandre Maros. - 2020.
xx, 70 f. il.

Orientadora: Jussara Marques de Almeida Gonçalves
Coorientador: Marisa Affonso Vasconcelos.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f.63-70.

1. Computação – Teses. 2. WhatsApp (Aplicativo de mensagens) – Teses. 3. Desinformação – Teses. 4. Disseminação da informação - Teses. I. Gonçalves, Jussara Marques de Almeida.II. Vasconcelos, Marisa Affonso. 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 Belkiz Inez Rezende Costa
CRB 6ª Região nº 1510

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

FOLHA DE APROVAÇÃO

Uma Análise de Mensagens de Áudio Compartilhadas em Grupos do WhatsApp

ALEXANDRE MAROS

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Jussara Marques de Almeida Gonçalves
PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES - Orientadora
Departamento de Ciência da Computação - UFMG

Marisa Affonso Vasconcelos
DRA. MARISA AFFONSO VASCONCELOS - Coorientadora
IBM Research

Ana Paula Couto da Silva
PROFA. ANA PAULA COUTO DA SILVA
Departamento de Ciência da Computação - UFMG

PROF. FABRÍCIO MURAI FERREIRA
Departamento de Ciência da Computação - UFMG

Humberto Marques Neto
PROF. HUMBERTO TORRES MARQUES NETO
Departamento de Ciência da Computação - PUC-MG

Belo Horizonte, 20 de Novembro de 2020.

"Tomorrow belongs to those who can hear it coming"
(David Bowie)

Resumo

O WhatsApp é um aplicativo de mensagens gratuito com mais de 1,5 bilhão de usuários ativos mensais que se tornou uma das principais plataformas de comunicação em muitos países, incluindo Alemanha, Malásia e Brasil. Além de permitir a troca direta de mensagens entre pares de usuários, o aplicativo também possibilita conversas em grupo, onde várias pessoas podem interagir entre si. Muitos estudos recentes têm mostrado que os grupos de WhatsApp desempenham um papel significativo como plataforma de disseminação de informações, especialmente durante eventos importantes de mobilização social. Nesta dissertação, complementamos esses estudos anteriores ao examinar o uso de mensagens de *áudio* em grupos de WhatsApp, um tipo de conteúdo que está se tornando cada vez mais importante na plataforma. Apresentamos uma metodologia para analisar mensagens de áudio compartilhadas em grupos publicamente acessíveis do WhatsApp, composta por várias etapas: (1) pré-processamento, (2) detecção de similaridade (para agrupar áudios com conteúdo equivalente), (3) reconhecimento de voz para transcrever os áudios, (4) detecção de desinformação, (5) categorização do tipo de áudio (para distinguir entre fala e música, assim como o gênero do locutor), (6) uma análise qualitativa com usuários voluntários e (7) análise de conteúdo e propagação.

Analisamos mais de 40 mil mensagens de áudio em seis meses, compartilhadas em 364 grupos. Primeiro, examinamos o conteúdo das mensagens de áudio fazendo uma análise de tópicos. Identificamos oito tópicos de discussão, quatro relacionados à política e contendo a maior fração de desinformação. Em seguida, extraímos características linguísticas psicológicas e identificamos que os áudios com desinformação tem uma presença maior de emoções negativas. Eles também costumam usar frases no tempo futuro e falam diretamente com o ouvinte usando palavras como “você”. Em contraste, estudos anteriores sobre desinformação em mensagens textuais compartilhadas no WhatsApp identificaram uma maior frequência do tempo presente e de termos para agregar a comunidade, como “nós”. A análise qualitativa mostrou que áudios com desinformação tendem a fazer o ouvinte sentir emoções negativas, como raiva. Os voluntários notaram que os áudios com desinformação tentaram dar crédito a suas

afirmações com fontes externas; no entanto, eles consideraram essas fontes como não confiáveis. O tom do locutor nos áudios com desinformação também foi considerado menos *amigável* e *natural* do que os áudios com conteúdo não verificado. Por fim, nossa análise de propagação mostrou que os áudios são compartilhados em intervalos curtos, com mais da metade deles sendo compartilhados em três horas, mas se espalhando mais lentamente do que o conteúdo textual e de imagem. Também descobrimos que os áudios contendo música costumam ser mais compartilhados do que apenas fala e tem uma vida mais longa. Além disso, os áudios com desinformação tendem a se espalhar mais rápido do que áudios com conteúdo não verificado e duram muito mais tempo na rede. Em suma, realizamos um estudo que, até onde sabemos, é o primeiro a abordar a comunicação de áudio em grupos de WhatsApp, demonstrando como analisar esse tipo de mídia, observando o conteúdo e a dinâmica de propagação, e comparando-a com outros tipos de mídia (texto e imagens), tipos de áudios distintos e áudios contendo desinformação. Nosso trabalho revelou que essa forma de comunicação segue padrões distintos de conteúdo de texto e imagem, principalmente no que diz respeito à desinformação, complementando assim a literatura.

Palavras-chave: WhatsApp, Comunicação por Áudio, Disseminação de Informação, Desinformação.

Abstract

WhatsApp is a free messaging app with more than 1.5 billion active monthly users that has become one of the leading communication platforms in many countries, including Germany, Malaysia, and Brazil. In addition to allowing the direct exchange of messages among pairs of users, the application also enables group conversations, where multiple people can interact with each other. Many recent studies have shown that WhatsApp groups play a significant role as an information dissemination platform, especially during important social mobilization events. In this thesis, we build upon those prior efforts by looking into the use of *audio* messages in WhatsApp groups, a type of content that is becoming increasingly important in the platform. We present a methodology to analyze audio messages shared in publicly accessible WhatsApp groups composed of several steps: (1) pre-processing, (2) similarity detection (to group audios with equivalent content), (3) speech recognition to transcribe the audios, (4) misinformation detection, (5) audio type categorization (to distinguish between speech and music, as well as by speaker's gender), (6) a qualitative analysis with volunteers users, and (7) content and propagation analysis.

We analyzed more than 40 thousand audio messages across six months shared in over 364 groups. We first looked into the content of the audio messages by doing a topic analysis. We identified eight topics of discussions, four related to politics, and containing the largest fraction of misinformation. We then extracted psychological linguistic features and identified that audios with misinformation had a higher presence of negative emotions. They also often used phrases in the future tense and talked directly to the listener by using words such as "you". In contrast, prior studies on misinformation in textual messages shared on Whatsapp identified a higher frequency of the present tense and terms to aggregate the community, such as "we". The qualitative analysis showed that audios with misinformation tend to make the listener feel negative emotions, such as anger. The volunteers noted that audios with misinformation tried to back their claims with sources; however, they often saw these sources as unreliable. The speaker's tone from the audios with misinformation was also considered less *friendly*

and *natural* than audios with unchecked content. Lastly, our propagation analysis showed that audios are re-shared within short intervals, with more than half of them being re-shared within three hours but spreading more slowly than textual and image content. We also found that audios containing music were often shared more than speech and had a longer lifetime. Moreover, audios with misinformation tend to spread quicker than unchecked content and last significantly longer in the network. In sum, we performed a study that, to our knowledge, is the first to tackle audio communication in WhatsApp groups, going over how to analyze this type of media, looking into content and propagation dynamics, and comparing it to other types of media (text and images), multiple audio types, and audios containing misinformation. Our work revealed that this form of communication follows different patterns from text and image content, especially when it comes to misinformation, thus complementing the literature.

Palavras-chave: WhatsApp, Audio Communication, Information Dissemination, Misinformation.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Mobile communication is currently centered around a few messaging apps, such as Facebook Messenger[1], Snapchat[2], Telegram[3], and WhatsApp[4]. Each of these options have specific advantages that make them more appealing to certain publics. Currently, one of the most used communication platforms in some countries such as Malaysia, Germany, Saudi Arabia, and Brazil is WhatsApp, which is a free world-wide messaging app created in 2009 that currently has more than 1.5 billion active monthly users [Iqbal, 2019]. The number of adopters of the platform is especially large in Brazil, where it is estimated that, in 2019, 116 million people, or 65% of the population, has access to it, and nearly everyone that uses a smartphone has WhatsApp installed in their cellphones [Ferreira, 2019].

WhatsApp has some key features that make it stand out. Firstly, the contents shared on WhatsApp are end-to-end encrypted, meaning that each user has a unique encryption key. The content is only encrypted and decrypted in the phones of those involved in the communication and cannot be seen by anyone else, not even the company responsible for that application. This feature is especially impressive compared to other popular messaging apps, such as Telegram and Facebook's Messenger, which do not have end-to-end encryption enabled by default. This encryption method dramatically increases the privacy of those involved in the conversation but also makes it difficult to track the dissemination of information at scale, which even has led to governments to propose breaking this encryption as a way to enable moderation and law enforcement on the platforms[5].

[1] https://www.messenger.com/.
[2] https://www.snapchat.com/.
[3] https://telegram.org/.
[4] https://www.whatsapp.com/.
[5] https://tecnoblog.net/274333/whatsapp-telegram-quebra-sigilo-proposta-cnj/.

Secondly, the platform offers a simple and easy-to-use set of features that allows anyone to quickly share texts, pictures, audios, videos, or files with individual users or several people at once, through the so-called groups. Anyone can create these groups, and the group administrators control access to it. Currently, there are two methods for joining them. The first one is by having the group administrator directly add a user. The second method is by having the group administrator generate a unique invitation link and share it with those interested in participating in the group. In the latter anyone who has access to the invitation link can automatically join the group. These links can be shared with a specific set of people, or they can be shared publicly, for example, in a social media post, thus making the group, from a practical perspective, *publicly accessible*. Individual groups are limited to 256 simultaneous members[6] but there are no limits on how many groups a person can create, invite, or participate[7]. End-to-end encryption is also enabled by default, and only those that are members of the group have access to the content of the messages.

Finally, there are tools for quickly spreading information in the app, such as broadcasting or forwarding. Broadcasting is a distinctive feature of WhatsApp, allowing users to create lists of users and groups. It allows one-to-many type of communication. When a user wants to send individual messages for each of these users, he can select this broadcast list. WhatsApp automatically submits the messages separately to each of these contacts. Currently, there is a limit on the number of contacts (individuals or groups) a message can be sent to at once[8], but the user can create multiple lists containing 256 groups or contacts and send a message separately to each one of these lists. The second method for quickly sharing information is forwarding, which allows users to forward one or more messages that they sent or received from someone (individually or through a group) to 5 different people or groups[9],[10] simultaneously.

1.1 Motivation

As a result of the high market penetration combined with the features described above, many recent events brought to light some worrisome behaviors emerging in the platform, more specifically those related to the dissemination of misinformation. As an example, in 2018, several men were killed by a mob after a rumor that they were kid-

[6] https://faq.whatsapp.com/web/chats/adding-and-removing-group-participants/.

[7] https://faq.whatsapp.com/general/requirements-for-broadcasting-a-message/

[8] https://faq.whatsapp.com/general/requirements-for-broadcasting-a-message.

[9] https://faq.whatsapp.com/general/coronavirus-product-changes/about-forwarding-limits.

[10] Before 2019, the forward limit was 20.

nappers went viral on WhatsApp in India [Meixler, 2018]. Moreover, in Brazil, it is estimated that at least 12 million people spread misinformation only in 2017 [Martins, 2018], while in the 2018 presidential election, this number grew even further. At the same time, there are reports on an increase in the number of people who use social media, including WhatsApp, as their primary source of news and information [de Assis, 2019]. These numbers, combined with the fact that the trust in the news coming from the Internet is decreasing, creates many discussions regarding the role of this kind of service in our society.

Several recent studies have analyzed content dissemination in publicly accessible WhatsApp groups [Resende et al., 2019b; Caetano et al., 2019; Resende et al., 2019a; Melo et al., 2019a,b; Bursztyn and Birnbaum, 2019]. These prior studies focused mostly on image and textual content, characterizing content properties as well as propagation dynamics and offering quantitative evidence of the use of the platform to spread misinformation (e.g., fake news) [Resende et al., 2019b; Caetano et al., 2019; Resende et al., 2019a; Reis et al., 2020]. These studies have shown that WhatsApp is not a mere communication tool but rather exhibits characteristics of social networks like Facebook, Youtube and Reddit, with the emergence of robust networks interconnecting users which facilitate the quick spread of information.

However, neither text nor image messaging can fully convey the sender's tone, urgency, emotion, or purpose as audio content can [Sherman et al., 2013], and some prior studies relied on audio media to capture these peculiarities [Ooi et al., 2014; Cunningham et al., 2020]. Moreover, audio communication is also a tool for digital inclusion, as people that have writing difficulties or that are illiterate can still use this tool. Indeed, it has been reported that the use of voice messages on WhatsApp has rapidly increased recently. In essence, over 200 million voice messages are sent by WhatsApp app every day in some regions[11]. Yet, despite the recent interest in WhatsApp by academia [Resende et al., 2019b; Caetano et al., 2019; Resende et al., 2019a; Melo et al., 2019a,b; Bursztyn and Birnbaum, 2019], no prior work has looked into the properties of audio content being shared in this platform.

1.2 Goals and Research Questions

Given that the spread of misinformation is increasing in WhatsApp groups and that more users are interacting with this type of content, whether it is just by idling lis-

[11] https://www.news.com.au/technology/gadgets/mobile-phones/why-people-are-switching-from-texting-to-voice-messages/news-story/d36d6d80cc0c71da168b4e8ec96924e7.

tening or forwarding to other contacts or groups, it becomes of interest to have a careful understanding of how these messages circulate in the app and how WhatsApp is exploited to allow and boost this kind of content. In this master thesis, we intend to investigate audio communication in political-oriented and public-accessible WhatsApp groups, searching for distinctive characteristics in terms of content properties and propagation dynamics. By doing so, we offer a first look into this form of communication in WhatsApp, aiming at complementing previous studies on textual and image content. In our investigation, we analyze audio messages carrying previously checked misinformation separately, aiming at identifying properties that distinguish them from the others (unchecked) audio messages. Given this scenario, we aim to answer the following research questions:

- **RQ1**: What are the characteristics of audio messages shared in publicly acessible WhatsApp groups in terms of content properties and propagation dynamics? How do they relate to prior findings for other types of content in that platform?
- **RQ2**: What are the introspect properties of audio content (e.g., the gender of speaker, music versus speech content) and how do these properties correlate with propagation dynamics?
- **RQ3**: How do the content and dynamics properties of audio messages carrying previously checked misinformation compare with the properties of the other audio messages?

To address these questions, we analyze a dataset obtained from Resende et al. [2019b,a] which consists of messages collected from *publicly accessible* and *politically-oriented* WhatsApp groups during two major social events in Brazil, the trucker strike from 21st of May and 2nd of July of 2018 and the presidential electoral campaign, from 16th of August to 28th of October of the same year.

1.3 Contributions

Unlike prior studies which focused on textual and image content, we here focus our study on *audio* content. Given the novelty of such effort, one of our key contributions is a pipeline of analysis, which consists of the following seven steps:

1. Pre-processing

The first step consists of a pre-processing phase to ensure that audios are in the same audio format, as the audios collected from WhatsApp can come in multiple formats, such as *.ogg* and *.wav* depending on how they are shared.

2. Similarity Detection

The next step consists of applying algorithms to detect similar audio content as a means to identify multiple instances of the same content, which is a required step to study how the same content is being spread across the platform. A simple bit-by-bit comparison is not a valid strategy in this case, as the audios shared in the groups have variations that make this approach not ideal. An audio file can be downloaded and re-uploaded to WhatsApp, which can cause the compression algorithm to take effect and change their representation. Other scenarios include multiple recordings of the same event (various people with their phones recording a speech with small change across recordings) or the recording of a recording, like someone recording the audio from an external speaker. Therefore, we opted for applying fingerprinting algorithms, such as Chromaprint [Bartsch and Wakefield, 2005; Jang et al., 2009; Porter, 2013], to detect audios that have the same content. Similar methods are used in many companies to do similar groupings with music, such as Shazam[12], a widely used app that can recognize millions of songs based on a few seconds of a recording.

3. Speech Recognition

The third step corresponds to applying speech recognition methods to transcribe the audio files using Google Cloud's Speech-to-Text[13] API. This allowed us to apply several natural language processing (NLP) algorithms to characterize the transcripted content.

4. Misinformation Detection

The fourth step is misinformation detection in the audios sent to the groups. As in Resende et al. [2019a,b], we relied on previously checked facts by various Brazilian fact checking agencies (such as Lupa[14] and Boatos[15]) and our transcripted audio messages. We consider an audio as containing misinformation when we find a match between the audio transcription and a previously checked news that was marked as containing misinformation from the fact checking agencies. Based on the news article and the transcription, we label audio messages as either carrying previously checked misinformation or not. Note that we cannot guarantee that the latter does not contain misinformation but rather only that it does not contain any misinformation that had

[12] https://www.shazam.com/.
[13] https://cloud.google.com/speech-to-text/.
[14] https://piaui.folha.uol.com.br/lupa/.
[15] http://boatos.org/.

been previously checked as so by the considered fact checkers. We thus refer to them as *unchecked* content.

5. Audio type categorization

The fifth step looks into a unique property that audio content has over only written text: the type of audio (spoken *versus* music) and the gender of the speaker. Using a pre-trained convolutional neural network described in [Doukhan et al., 2018], we were able to identify the gender of the speaker for each audio collected with an average F1 score of 0.97 over our test set. We dive into gender analysis as several studies have shown that gender matter is an important factor in user behavior [Lorigo et al., 2006; Tang et al., 2011; Mueller et al., 2017] and we can accurately identify the gender of the speaker based on an audio message. Exploring this categorization we were able to determine that whereas there is no clear correlation between gender and audio message propagation dynamics, music (as opposed to spoken content) tends to be much more shared.

6. Qualitative Analysis

The sixth step was conducting a qualitative analysis of the audios with two phases, an interview and a survey. The main objective of this analysis was to gather information about how the users interact with WhatsApp groups and audio messages (e.g., how many groups they participate in, how many audio messages they receive), and gather insights on the differences that audio with misinformation has over unchecked content. To do so, we presented multiple audios with both types of content and collected information about their perception. By doing so, we noticed specific reactions to certain audios, such as anger towards fake messages.

7. Content and Propagation Analysis

Finally, the seventh step corresponds to analyzing content and propagation properties. Based on the the information gathered in the previous steps, we analyzed the general topics of discussion and psychological and linguistic features of these transcripted audios. We then looked into how these audio messages propagate in the WhatsApp groups, looking at characteristics such as total duration (lifetime) and time between consecutive shares of the same content. Finally compared how these characteristics differ based on different audio types, such as misinformation versus unchecked, gender and speech versus music.

Our main findings revealed that audio communication is widely used in the 364 monitored WhatsApp groups, with more than 40 thousand audio messages in the app across six months. Also, audios are often re-shared within reasonably short intervals: 60% of audios are re-shared within 3 hours and 20% within 3 minutes. Audio messages also spread more slowly and remain for shorter periods compared to textual and image content, which could reflect the larger effort required to listen to an audio message compared to a text. Based on the misinformation detection, we marked over 120 unique audios that were shared more than 2000 times across 260 groups during the monitored period. We observed that audios with misinformation appear in more groups and are shared by more users than their counterparts. Audios that contained music had a higher reach than speech audios, being shared by 28% more users and reaching 43% more groups. In the election period, these audios were often campaign jingles. We found no evidence that gender affected the number of times an audio message was shared. Lastly, we noticed many particular characteristics that emerged more often in audios with misinformation, such as a call to action (actively asking the listener to take some action, such as share the audio) and being more related to negative emotions.

1.4 Outline

The remainder of this thesis is organized as follows. In Chapter 2, we present background information and discuss prior studies closely related to our present effort. In Chapter 3, we present the methodology we adopted to analyze audio messages in WhatsApp, from collecting the data, treating them, grouping similar audios, and transcribing them. We also present an overview of the collected dataset. In Chapter 4 we present the main findings from our analysis, including a characterization of the content and propagation properties of all audio messages as well as of audio messages carrying previously checked misinformation. Finally, Chapter 5 summarizes the dissertation, offering conclusions and some directions for future work.

Chapter 2

Background and Related Work

In this chapter, we go over the background and related work for this thesis. In Section 2.1, we review some key concepts and algorithms related to our work, notably audio similarity detection, speech recognition, and text representation. In Section 2.2, we discuss prior studies that served as basis and a starting point for this thesis, including prior studios on audio analysis, misinformation outside of WhatsApp, and studies that are directly related to WhatsApp. Finally, in Section 2.3, we present a summary of this chapter.

2.1 Main Concepts

2.1.1 Audio Similarity Detection

One of the main steps in our analysis consists of grouping similar audios. This is not a trivial task to do since many audio variations exist, and a simple bit-level comparison would fail to correctly group many similar files. These differences are caused by many factors, such as (i) different audio compression algorithms and different audio formats (e.g., .mp3 and .wav); (ii) it can be a recording of a recording (a person may record someone else's priorly recorded audio); (iii) the audio can be cut a few second short or have a few seconds more than the original.

There are alternative approaches to compare similar audios. One approach is via fingerprinting [Cano et al., 2002]. This process consists of taking an audio file as an input and transforming it into uniquely identifiable features. Many existing algorithms implement this approach. Examples are Chromaprint[1] which is an implementation of

[1] https://acoustid.org/chromaprint

an algorithm inspired by Bartsch and Wakefield [2005]; Jang et al. [2009], Landmark[2], a Matlab implementation of Wang et al. [2003], and Echoprint [Ellis et al., 2011].

Chromaprint, which we use in this thesis, converts the audio signal into the frequency domain by performing short-time Fourier transformations. The resulting spectrum is converted to 12 bins representing the chroma of the signal. Each bin represents one of the 12 notes of the diatonic scale. After this, filter shapes are used, and a 12-by-16 sliding window is moved over the chromagram one sample (a small-time interval) at a time, which sums the amount of energy in the sample. Each filter quantizes the energy value to a 2-bit number, which is subsequently combined and converted to a 32-bit integer. The comparison between these fingerprints can be calculated, producing a number between 0 and 1 (1 being the perfect match). This process has been widely used and validated, and several companies such as Shazam and Spotify use variations of this algorithm [Bartsch and Wakefield, 2005; Jang et al., 2009; Porter, 2013].

Another approach is by calculating the distance between different Mel-Frequency Cepstral Coefficients (MFCC) [Jensen et al., 2006]. MFCC's are also a high-level representation, where the audio is converted into the frequency spectrum with a Fourier transform, and several cosine transformations are applied. The MFCC is represented as a time-series of these high-level representations. The algorithm Dynamic Time Warping [Müller, 2007] is used to compare different audio representations. This algorithm measures the similarity between two temporal sequences. This process is widely documented in the literature [Berndt and Clifford, 1994; Logan and Salomon, 2001; McKinney and Breebaart, 2003]. A drawback of this approach is the time complexity of the algorithm, which is remarkably higher than the fingerprinting methods. There are approximate ways to calculate the difference; one example is shown in Salvador and Chan [2007]. However, the longer the audio is, the higher the error, and the numbers produced to represent the difference is not easily comparable. For this reason, we here adopt the aforementioned Chromaprint method which produces similar results with superior time performance.

2.1.2 Speech Recognition

Speech recognition is the task of converting audio and speech automatically into text. This is a complicated task in the sense that there is no way of quickly creating algorithms to transcribe audio-encoded speech into text. It is also problematic, as there are many languages in existence, with different nuances and accents based on the region the speaker is from.

[2] https://www.ee.columbia.edu/~dpwe/resources/matlab/fingerprint/

Picone [1993] presented a few techniques used in the early days of speech recognition. The primary strategy used is composed of two steps, namely signal modeling and network searching. The first step consists of converting sequences of speech samples to vectors that represent events (e.g., words) in a probability space, and the second one is the task of finding the most probable series of events given syntactic constraints. Juang and Rabiner [1991] described a second method by using Hidden Markov Models (HMM) to model speech and train the probabilities given the audio signal.

These early techniques, even though they were powerful at times, often did not perform well in the day-to-day audio transcriptions, where noise and accents were often commonly found. With the popularization of deep learning methods, a new approach to solve the problem of audio transcription emerged. Given a raw audio file x and their respective transcription y, neural networks were applied to find rules to transcribe these audios automatically. Saon et al. [2017]; Lüscher et al. [2019] used a Long short-term memory (LSTM) network, which is a Recurrent Neural Network (RNN) designed to handle ordered sequences of data, such as audio streams. Other approaches also use Deep Neural Networks (DNN), as the method proposed in [Han et al., 2019], based on the transformer model on a self-attention mechanism.

Typically, DNN requires a vast amount of labeled data to train efficiently. This limits the option of open-source tools, as often the datasets are not publicly available (or are centered mostly around the English language). When it comes to transcriptions of Brazilian Portuguese audio, the literature and tools available are quite limited for that reason. However, many private companies who have research in cognitive services provide services that transcribe audios to users, and they often use DNN to do so. Examples of companies that offer speech recognition engines to users are Microsoft[3], IBM[4] and Google[5]. Quintanilha [2017]; Herchonvicz et al. [2019] analyzed these three cloud-based speech recognition engines for the Brazilian Portuguese language and concluded that the Google Cloud Speech-to-Text API had a considerably lower error rate. This is why we adopted this method in our work.

2.1.3 Text Representation

Humans can efficiently process text and understand sentences, images, audio, and videos. However, we use different representations to ease computer interpretation. When discussing these representations, we often use the concepts of *documents* and *terms* do describe them. A document d ($d \in D$, where D is a collection of documents)

[3] https://azure.microsoft.com/services/cognitive-services/speech-to-text/
[4] https://www.ibm.com/cloud/watson-speech-to-text
[5] https://cloud.google.com/speech-to-text

is represented as a collection of terms t, that can be words, stems, phrases, or any other unit derived from the text of the document. One common way to represent textual sentences is vector representations, with one position for each term, and a weight representing this term's relevance to the sentence. Each of these representations has pros and cons, and their usage varies according to the particular goal at hand. The most common approaches for creating these representations are one-hot encoding, bag-of-words, TF-IDF, and word-embeddings.

A one-hot-encoding is a representation of documents into binary vectors. A document d is represented as a binary vector v of size V, where V is the vocabulary size of the whole collection. For each term t in document d, the corresponding position of the term t on vector v is marked as one. The one-hot-encoding is one of the simplest ways of representing a vocabulary, but it suffers from a few drawbacks, such as not considering the word order and the high sparsity of the vector. Bag-of-words is an evolution of the one-hot-encoding with term frequencies. Each document is still represented by a vector of size V, but instead of having a binary representation simply indicating the presence or not of a word w, we have the number of times w appeared in the document. It still has many of the drawbacks that one-hot-encoding has [Zhang et al., 2010].

The term frequency gives more information on a document than the one-hot-encoding method; however, it does not contain information about the frequency of a word, relative to the other words. Therefore, words that are very common in a language (e.g., "the", "and", "but" on English) appear frequently and, therefore, have a high term frequency, but do not describe the document well. To solve this problem, Salton and Buckley [1988] created the term frequency-inverse document frequency (TF-IDF) metric, which takes into account weighting factors. This representation is a product of two metrics, term frequency and inverse document frequency. The term frequency $tf(t, d)$ is the number of times t appears in a document d. The inverse document frequency, $idf(t, d, D)$, is a measure of how common or rare the term t is across all documents in the collection D and, thus, how discriminative the word is of this document with respect to the others. If the word is used in many documents it does not carry much discriminative information. The complete formula for the $tf - idf$ representation is given as follows:

$$idf(t, D) = \log \frac{N}{\{|d \in D : t \in d|\}} \tag{2.1}$$

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \tag{2.2}$$

where N is the number of documents and $\{|d \in D : t \in d|\}$ represents the number of

documents where the term t appears. The tf-idf gives us the importance of a word in each document in terms of both descriptive (TF) and discriminative (IDF) capacities. However, the same problems as those associated with one-hot-encoding and bag-of-words such as high dimensionality and sparsity with no semantic relationship between words.

In contrast, word embeddings are a recent attempt to transform words into a low dimensional vector space that incorporates the meaning and semantics of the words. The main idea is that each position of the space corresponds to a latent feature in the word. Words with a similar meaning (e.g., "fruit" and "pineapple") are close together in this space [Pennington et al., 2014; Goldberg and Levy, 2014]. These representations are continuous, low dimensional, dense vectors that incorporate meaning, and semantic content. This technique has been widely studied and applied in several tasks, including sentiment analysis [Giatsoglou et al., 2017]. However, for representing documents, it is required to aggregate each word's representation. Common aggregation techniques are coordinate-wise mean, min, max, or even the concatenation of every representation.

In this study, we adopt a bag-of-word representation for topic modeling and TF-IDF for similarity classification between two documents to detect misinformation in the audios. The use of word-embeddings to perform sentiment analysis and other similar analyses that use this representation were left as future work.

2.2 Prior Related Studies

2.2.1 Audio Analysis

Earlier studies on audio analysis mostly focused on only analyzing the textual output provided by speech recognition systems [Larson et al., 2012b,a] to index and retrieve audio files as well as try to capture the emotion of the content. More recent studies applied machine learning algorithms to more accurately predict the emotion of the speaker by using the audio file alone [Ooi et al., 2014; Cunningham et al., 2020]), while some explored out multi-modal approaches, using not only the audio but also text [Yoon et al., 2018] and, if available, even video [Ortega et al., 2019] to predict the emotion of the sentences.

Kotti and Kotropoulos [2008] explored more than 1300 features such as loudness, mel-frequency cepstral coefficients (MFCCs), frequency, alongside a support vector machine (SVM) to classify the gender and sentiment of the speaker. They achieved a high accuracy (98%). However, the model was only tested in acted and noiseless audios and not on real-life scenarios. Meinedo and Trancoso [2010] tackled both gender and

age classification, proposing a similar approach by extracting features from the audios and applying several classification models such as Gaussian Mixture Models (GMM) and SVMs. To approximate the age of the person speaking, they separated their target into four groups: child, young, adult, and senior. The model achieved a higher accuracy for gender (95% in the best case) but low accuracy for age (56% in the best case)

With the rise of deep neural networks, Doukhan et al. [2018] tackled the gender prediction problem with Convolutional Neural Networks (CNN) with the help of Gaussian Mixture Models and i-vectors. They trained the network using an internal database of 2284 French speakers and obtained an F-measure of 96%. The proposed system was also designed to handle long audio files and multiple speakers by processing segments of the audio individually. The proposed classification technique can also differentiate between a spoken segment and a music segment, thus allowing to detect intervals where music is played [Doukhan et al., 2018].

In Yang et al. [2019], the authors explored features such as energy, humor, and creativity and use them to predict the *seriousness* and the popularity of podcasts. Prior studies on this topic relied only on automatic speech recognition, ignoring vocal, musical, and conversational properties (energy, creativity). They proposed an adversarial learning-based approach to generate features for popularity prediction which outperformed the current state-of-the-art hand features.

Building on those prior studies, we here focus not only on textual information generated by speech recognition, but also on characteristics that are specific to audio, such as the gender of the speaker and audio type classification (speech and music).

2.2.2 Misinformation Analysis

There is extensive literature centered around misinformation on the Internet, particularly in social media. We here go over only a few of these studies, focusing on a few works on misinformation detection and misinformation analysis in the past few years across the most popular social media or revolving around similar events as this thesis, such as elections. Characterizing fake news or content with misinformation is not an easy task. Some automatic approaches have been already proposed, like, for example, in Qazvinian et al. [2011], the authors proposed information retrieval techniques to find tweets with misinformation. Despite good results for the training set, the proposed methods exhibited noisy results for the test set. In Conti et al. [2017], the authors argued for the difficulty of classifying misinformation utilizing only propagation characteristics on a Facebook network. Lastly, Kumar and Geethakumari [2014] proposed a model based on cognitive psychology to identify misinformation in messages.

The analyzed factors were coherence, credibility, consistency, and general acceptance of the message. The tool would alert the user in case they detected misinformation. None of the methods showed high precision, being used only to alert users or monitor certain behaviors. Many authors, such as Resende et al. [2019b,a] use external agencies to identify content containing previously checked misinformation; however, this is a costly process, as requires interaction from many people, such as journalists, to classify misinformation.

In Fourney et al. [2017], the authors studied fake news propagation in the preceding months of the 2016 United States presidential election. The study verified the existence of misinformation on several websites. It showed that there is a correlation between the proportion of votes for the Republican candidate Donald Trump in each state versus the fraction of users who accessed websites that contained misinformation.

Bessi and Ferrara [2016] cited Twitter as a misinformation source that can negatively affect democratic political discussion. The authors observed the significant influence and control that bots have over a discussion. Out of almost 3 million distinct users involved in political discussions, 400 thousand were bots and were responsible for 3.8 million tweets (the equivalent of one-fifth of all collected tweets). These numbers are worrisome since these bots can act in an orchestrated way to influence and promote discussion, impulsing content with misinformation, and influencing what is being discussed by real users [Allcott and Gentzkow, 2017]. The bots are not only targeting politics, but also several other areas, such as debates regarding vaccination campaigns [Broniatowski et al., 2018], and are not limited to Twitter, but are also present in other social networks, such as Facebook and Reddit [Ferrara et al., 2016].

2.2.3 Studies on WhatsApp

The great popularity of WhatsApp in many countries and its rise to become one of the most used messaging apps on them made the platform stand out from the rest. The central role that the application had over major social events, like the Brazilian 2018 election [Martins, 2018; Loubak and Achilles, 2019], were important factors to trigger many studies to understand the aspects of this platform.

Tardáguila et al. [2018] points out concerns regarding the easiness of fake news propagation in groups of the app, and that there are some tools to facilitate the spread of this news. The content is created and shared with some activists, which then forward the information in their groups. The authors show that from a sample of 50 most shared images collected from publicly accessible WhatsApp groups, 56% of them had some misinformation. The authors also discussed some strategies to slow down the

propagation of this type of content such as, restricting message forwarding, restricting the broadcast tool, and limiting the size of new groups.

More recently, many studies have analyzed user behavior and content dissemination in publicly open WhatsApp groups. Melo et al. [2019a] proposed a general data collection methodology. This methodology consists of monitoring these groups with the help of external mobile devices and downloading daily the data shared within each group, persisting this data in a database. Several attributes are extracted, such as group ID, user ID, timestamp, message, audio. The data is anonymized so that individual users are not recognizable. All this data is exposed to a website in which the people involved in the project (journalists, researchers) can access it [Melo et al., 2019a]. Seufert et al. [2016] analyzed the implications of group conversations on mobile network traffic, such as the usage of WhatsApp and the distribution of users per group. Based on an interview, they consulted 200 people located in Germany on their WhatsApp usage history. They found out that the groups analyzed in their study were limited to few members (on average nine) and contained only people they were closely related to, which varies vastly from what was found in this thesis, showing how the usage of WhatsApp can change across the years or regions. Finally, through the analysis of message histories, the authors compiled a communication model based on a semi-Markov process. The model lists the probability that a text post is followed by another text post or a media post (image, video, or audio) and the probability that a media post is followed by a text post or another media post. With this model, they can simulate messages sent in the groups. The authors left, as a future work, breaking the media posts into their respective subcategories [Seufert et al., 2016].

Caetano et al. [2019], in turn, proposed a hierarchical methodology to analyze user interactions on publicly accessible WhatsApp groups by using a cascade model. They analyzed cascades associated with more than 1.7 million messages posted in 120 groups over one year while also looking into differences between misinformation and other content shared. They analyzed the structural and dynamic properties of cascades in political and non-political groups distinguishing between cascades carrying content with misinformation from the rest. One key observation is that cascades with misinformation tend to be deeper, reach more users, and last longer in political groups than in non-political groups.

In Bursztyn and Birnbaum [2019], the authors analyzed differences between WhatsApp groups that are primarily left-wing and right-wing. The authors found that right-wing groups are more tightly connected and geographically distributed, while also sharing more multimedia messages. Melo et al. [2019b] analyzed the speed with which information is spread in WhatsApp groups considering an epidemiological model. The

authors identified parameters, that could control, difficult, or slow down the propagation of misinformation, such as how many people can a person forward a single message [Melo et al., 2019b].

Lastly, this thesis directly complements the following two works. In Resende et al. [2019b], the authors studied the types of content shared in publicly accessible WhatsApp groups during two periods of 2018 in Brazil, namely, the trucker strike and the presidential election. To do so, the authors submitted a search query across Google, Twitter, and Facebook search engines looking for a WhatsApp invite link so they could join the groups. They restricted the search space by including in each query a word from a dictionary related to the 2018 Brazilian elections, such as names of politicians, political parties, and words associated with political extremism. The study initially describes the amount and type of content shared across the groups (text, images, videos, and audios). The authors also created models of the group network, concluding that WhatsApp is much more than a simple communication tool with features and behaviors similar to other social networks, promoting ways for information to become viral. After that, the authors deepen their study by focusing on the content and propagation properties of messages carrying images. They proposed a method to identify misinformation in images shared across the groups. They concluded that images with previously checked misinformation spread more quickly than messages with previously unchecked content. Finally, they looked into the interplay between WhatsApp and other platforms, finding that the vast majority (95%) of unchecked images were first posted on the Web and then in WhatsApp, and in contrast, only 45% of images with misinformation were shared first on the Web.

In Resende et al. [2019a], the same authors extended their prior work by focusing on textual content. They analyzed textual misinformation shared in WhatsApp groups and compared how different attributes such as message size, psychological linguistic features, topics, sentiment, and frequent terms vary in comparison to unchecked messages. Lastly, they measured the propagation dynamics of text messages across the groups. They concluded that messages with misinformation are generally shared more times and more quickly than unchecked messages.

2.3 Summary

Our present effort focuses on analyzing audio messages on WhatsApp groups. To the best of our knowledge, there has not yet been a detailed study of how audio communication is used in WhatsApp. This thesis adds to the current literature by providing a

way to analyze audio content in messaging apps and extract meaningful characteristics related to audio content, such as the gender of the speaker. In addition, we explore content and propagation dynamics of audios across publicly accessible groups while looking at how the presence of misinformation affects these attributes. We also add to previous studies on WhatsApp by focusing on a different type of content – audio – and contrasting our findings with those previously obtained for textual and image content.

Chapter 3

Methodology

In this chapter, we describe the dataset collection process used in this work, as well as the methods used to pre-process, group audios with similar content together, transcribe them, detect content with misinformation and classifying the type of audio between music and speech, and when they are speech we also classify the speaker's gender. Finally, we overview some important characteristics of the dataset, such as the amount of audio shared and audio duration distribution.

3.1 Main Steps

We start by outlining the main steps of the methodology adopted for building the dataset, first discussing the data collection, how we grouped audios with similar content and then approaching the speech recognition problem for a deeper dive into the content being shared.

3.1.1 WhatsApp Dataset Collection

The dataset used here was previously collected by Resende et al. [2019b]. This dataset collection was performed in two steps. Firstly, the authors looked into search engines from Facebook, Twitter, and Google for the URL pattern "chat.whatsapp.com", which is generated when a group admin creates an invitation link: by clicking on the link, a person is automatically admitted into the group. The authors restricted the search to only capture groups that could be related to Brazilian politics by including in the query keywords and terms related to political terms. These keywords include names of politicians, political parties as well as majorly discussed themes (e.g., Marielle Lives, Lei Rouanet, Feminism, Patriot). After entering these groups, a manual analysis took

place, in which the authors removed groups that were unrelated to politics or had broken links.

For the second phase of the dataset collection, the authors selected 364 groups for monitoring. This number is lower than the complete list of possible groups due to the memory limitations on the devices used for the collection. A cellphone with a valid phone number is required to monitor each group; therefore, the maximum number of groups that could be monitored was limited by the resources available. After joining each monitored group, all messages shared in each group were collected via an API used by Melo et al. [2019a] called *WebWhatsapp-Wrapper*[1] which uses Python to receive messages by *WhatsApp Web*, a web-version of WhatsApp. Messages carrying content in all different media types were collected, such as text, images, videos, and audio. For each message, the timestamp, group identification, and user identification were collected. All user information was anonymized to protect the privacy of those involved. The data collection took place during 6 months in Brazil: from April 16th to October 28th 2018[2].

Specifically, for the audio content part of the process, different audio formats were identified when collecting the data from the API provided, such as *ogg* and *mp3*. For this reason, an initial pre-processing step is required to convert all of the audio formats into a single format. We converted all the files to the Free Lossless Audio Codec (FLAC) format using an open-source Python library called *pyDub*[3].

3.1.2 Grouping Audios with Similar Content

As discussed in detail in Section 2.1.1, the identification of audios with similar contents cannot be performed by merely comparing whether both files are identical (on a bit level) as there might be many variations of the same audio. For example, different compression methods could be used; one audio file could be cut slightly shorter, or different devices could record the same audio. Thus, to identify and group audios with the same content, we employed a fingerprinting method.

Specifically, we used an open-source library called *Chromaprint*[4], which processes and transforms the audio frequency in musical notes and uses this new representation to compare different files. This method is used in many other studies, such as Bartsch and Wakefield, 2005; Jang et al., 2009; Porter, 2013; Bhatia et al., 2018]. The method

[1] https://github.com/mukulhase/WebWhatsapp-Wrapper.
[2] We thank the authors for sharing the data with us.
[3] https://github.com/jiaaro/pydub.
[4] https://acoustid.org/chromaprint.

returns a score between 0 and 1, where 0 indicates completely different audios, and 1 represents a perfect match.

We only compared pairs of audios with durations that differ by no more than two seconds from each other as more significant differences were assumed to reflect different contents. For each pair of audios a_1 and a_2, the *Chromaprint* algorithm calculated their respective audio fingerprints, and the similarity score was calculated between them. We stored all of this comparison in a database to manually analyze them and select the score threshold γ that would minimize false positives to avoid grouping audios with different contents together. To find the desired γ threshold to apply to our grouping method, we manually analyzed a set of audios given different ranges of γ. To do so, we defined four ranges of possible values ($0.95 \leq \gamma < 1.0$, $0.90 \leq \gamma < 0.95$, $0.85 \leq \gamma < 0.90$ and $0.80 \leq \gamma < 0.85$) and selected 50 random pairs of audio that fall into each of those similarity ranges (e.g., selected 50 random pairs of audios with γ above or equal 0.95, 50 random pairs of audios with γ bellow 0.95 and above or equal 0.90, and so on). Then, we manually analyzed this sample (i.e., listened to both audios and decided whether they had the same content) to check the precision achieved in each threshold. The results are shown in the following table:

Table 3.1. Precision of similar audios with different thresholds

Threshold γ	True Positive	False Positive	Precision
$0.95 \leq \gamma < 1.0$	50	0	1.0
$0.90 \leq \gamma < 0.95$	50	0	1.0
$0.85 \leq \gamma < 0.90$	44	6	0.88
$0.80 \leq \gamma < 0.85$	37	13	0.74

Table 3.1 shows the results of this comparison by presenting four different ranges of γ threshold, the number of audio pairs for which both manual evaluation and the fingerprint comparison agreed they have similar content (*true positive*) and the number of audio pairs that were considered as similar content by Chromaprint though the manual evaluation indicated they were indeed different (*false positive*). For each range, the table also shows the precision associated with the Chromaprint algorithm defined as the ratio of the true positives to the sum of true and false positives.

With γ lower than 0.9, we started identifying some false positives, for instance, people speaking softly or audios composed majorly of loud noises, such as sirens. We concluded that all comparisons with a score of 0.9 or above indeed consisted of the same content. Lower values otherwise resulted in some false positives. Thus, we selected the threshold $\gamma = 0.9$ and used it to group the audios with similar content. Each audio pair that had the similarity score equal or above this threshold was merged together,

keeping track of their respective timestamps, groups in which they were shared, and the users who shared them. We randomly chose an audio file as representative of each group so we could apply the next steps of the processing pipeline, such as the transcription, to only one of these examples, and not repeat the process over all (near) duplicates of the same content.

3.1.3 Audio Transcription

As the second step of our methodology, we also performed a speech-to-text transcription of the audio messages in our dataset. To that end, we employed an automatic translation method, namely the Google Cloud's Speech-to-Text[5] API, presented in Section 2.1.2. We explored other publicly available tools for speech recognition; however, most of the open-source pre-trained neural networks available were trained for the English language, and no viable option was found for the Portuguese language aside from commercially available tools, such as Google's, Microsoft's Azure and IBM's. In recent studies, Quintanilha [2017] compared these three tools and observed that the API provided by Google had a considerably lower Word Error Rate (WER) and Label Error Rate (LER). This is why we chose it.

The Google Speech's API receives as input an audio file and returns the transcription t and a score s, $0 \leq s \leq 1$, which reflects the confidence of the model on the transcription produced. We validated the quality of the transcription by asking 35 volunteers to judge the transcriptions performed on a sample of 300 audios. Specifically, each volunteer was asked to first listen to the audio and then respond to whether the transcription correctly reflected the audio content. Each volunteer responded using a 0-4 Likert scale, where 0 indicates complete disagreement and 4 indicates complete agreement. Three volunteers judged each audio.

Figure 3.1, shows a scatter plot where each point represents an audio from the selected sample. For each audio, the graph shows the Google speech confidence scores versus the score given by the volunteer. Since each audio was linked to three different volunteers, we averaged the score for each audio. We found a strong correlation between the transcription scores and the average responses of the volunteers, with a Pearson correlation coefficient equal to 0.86. Indeed, we found that all audios with transcription scores above 0.8 received, on average, 3.6 points by the volunteers, suggesting high transcription quality. In turn, audios with scores bellow 0.8 received, on average, 1.7 points, suggesting very poor transcriptions (on average). Interestingly we found that almost all audio messages containing music content fell in the poor transcription

[5] https://cloud.google.com/speech-to-text/.

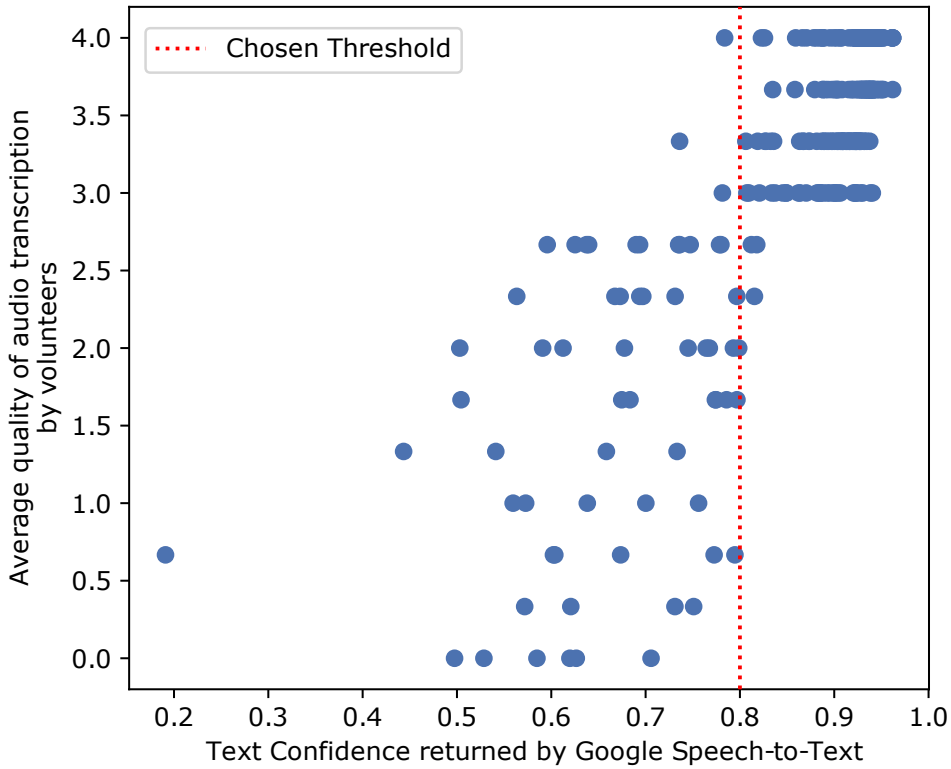


Figure 3.1. Google Confidence score on the transcription *versus* manually evaluated score

category, possibly due to the presence of melody in them. Thus, for analysis that considers the audio messages' content, we look at only the audios that presented a transcription score of at least 0.8, which corresponds to 70.9% of the audios.

3.2 Misinformation Detection

Detecting misinformation is a challenging task. Many authors proposed different methods for detecting misinformation, such as using a fuzzy analytic hierarchy process to assign weights to some proposed metrics [Baeth and Aktas, 2019], or by detecting social bots as an initial step for computation fact-checking [Menczer, 2016; Antoniadis et al., 2015]. Another approach is by relying on fact-checking journalists and agencies, where they specialize in assessing the truth of a public claim by seeking reliable claims and analyzing facts, images, and videos and directly contacting those involved in these claims [Graves, 2013].

We opted to utilize fact-checking agencies to find misinformation in the content of our analyzed audios. In the past few years, many of these journalists and groups

created several news portals where pieces of information shared online in different media formats, such as text and audios, were fact-checked, and the results were made publicly available. We utilized a dataset collected by Resende et al. [2019a] containing information fact-checked, which were scraped via a Python API. This dataset consists of a list of fact checked claims collected from the following fact-checking agencies in Brazil:

- Aos Fatos[6]: A fact-checking platform composed by independent journalists.
- G1[7]: A Brazilian news portal maintained by the Grupo Globo.
- E-farsas[8]: A website created in 2002 to disprove rumors that appeared on the Internet.
- boatos.org[9]: An independent news website created to compile and check news with potential misinformation .
- Veja[10]: A Brazilian magazine and news portal that has a special column for fact-check news.
- Agência Lupa[11]: A news agency specialized in fact-checking information, a member of the International Fact-Checking Network[12].

We then calculated the similarity of each audio transcription A of our audio dataset obtained in Section 3.1.3 with a fact-checked news B, marked as containing misinformation, from these fact-checking agencies. We computed the cosine similarity based on the TF-IDF vectors of the audio transcriptions with all of the collected fact-checked articles in the dataset. We applied a pre-processing step on both the audio transcription and the fact-checked news, removing stop words, and using lemmatization.

As seen in Section 2.1.3, TF-IDF (term frequency-inverse document frequency) [Salton and Buckley, 1988] is an approach to numerically represent text. In this approach, we have a collection of documents consisting of all the transcripted audio messages and the fact-checked news articles marked as containing misinformation. Each

[6] www.aosfatos.org.
[7] www.g1.globo.com.
[8] www.e-farsas.com.
[9] www.boatos.org.
[10] www.veja.abril.com.br/blog/me-engana-que-eu-posto/.
[11] www.piaui.folha.uol.com.br.
[12] https://piaui.folha.uol.com.br/lupa/2020/04/02/lupa-selo-ifcn/.

document in the collection is composed of a TF-IDF vector v, that has size V, which is the length of the vocabulary size of the collection (transcripted audio messages and fact-checked articles). Each position i of vector v is linked to a specific word in the vocabulary and represents their respective TF-IDF value.

Given A and B the TF-IDF vectors representing two texts (e.g., an audio transcription A and previously checked as fake claim B, the cosine similarity of A and B is given by:

$$similarity = \cos(A, B) = \frac{AB}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2}\sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{3.1}$$

After that, we manually analyzed the 300 pairs of texts (audio transcription and previously checked as fake claims) with the highest cosine similarity. Our goal with this analysis was to check whether the audio transcription contained the same content as the previously checked fake claim. We found that only 100 out of the 300 transcriptions analyzed indeed carried the same content as the claim they were matched to with the highest similarity according to the cosine similarity. These audios were marked as containing misinformation for further analyses. All other 200 audio messages, as well as all other audios with less similar content compared to the collected claims, were marked as *unchecked*[13].

3.3 Speaker gender detection

One of the unique characteristics that we can observe only in audio messages is the audio category specified as either music or speech. If the audio is speech, we can also estimate the gender of the speaker.

To achieve this task, we used a framework presented in Doukhan et al. [2018], which was designed to perform large-scale gender equality studies based on men and women speech-time percentage estimation. We tried using pyAudioAnalysis[14] but it did not perform as well on our dataset. The tool provided by Doukhan et al. [2018] splits audio signals into zones of speech, music and noise. Speech zones are then split again into two segments, the speaker gender (male or female in this case). Zones corresponding to speech over music or speech over noise are classified as speech.

The framework works by firstly extracting features based on a 25ms sliding window with a 10ms shift which are then directly fed it into a convolutional neural network

[13] We use the term unchecked to emphasize that all we can state is that they are not similar to any previously checked as fake claim collected. Thus they may or may not carry misinformation.

[14] https://github.com/tyiannak/pyAudioAnalysis.

(CNN). The CNN has 15 hidden layers and it is used to classify each window into each category. To train the model, they used the INA's Speaker Dictionary [Salmon and Vallet, 2014; Vallet et al., 2016], which is one of the largest manually annotated speaker database. It consists of 32000 french speech excerpts, corresponding to 1780 male (94 hours) and 494 female speakers (27 hours). To evaluate the model, they used the REPERE challenge corpus [Giraudel et al., 2012], which contains French TV streams and obtained an accuracy of 97.4%.

They made the pre-trained model available through an open-source library named inaSpeechSegmenter[15], which we here use for this classification task. Therefore, we classify Whatsapp audio messages into three categories: Speech, Music, and random noises/inactivity periods[16]. For audio messages classified as speech audio, we also extracted the gender of the speaker. Audios messages that contained both male and female speakers were classified as the predominant gender, that is, the gender of whoever spoke for the longest time.

Since the original model was trained for the French language, we had to verify its performance on our dataset composed of Brazilian Portuguese audios. To evaluate the model classification efficiency in our dataset, we asked 25 volunteers to manually classify a sample of our dataset's audio files. We randomly selected 300 audios and asked them to classify each audio in their respective categories: speech (and if it was speech, which gender was predominant) or music. Three different people annotated each audio file.

We measured the inter-annotator agreement using Fleiss's kappa coefficient (κ) [Fleiss, 1971]. Fleiss's kappa coefficient (κ) is a statistic to measure the agreement between annotators for categorical items. The metric can be interpreted as the extent to which the observed amount of agreement exceeds what would be expected if all ratings or classifications were completely random. Its result can be interpreted as follows: values ≤ 0 as indicating poor agreement, 0.01 to 0.20 as slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, 0.61 to 0.80 as substantial, and 0.81 to 1.00 as almost perfect agreement. The Fleiss's kappa coefficient (κ) [Fleiss, 1971] obtained was 0.86, indicating almost perfect agreement. Only 12 of the 300 audios had some disagreement, in which case we manually reviewed them. With that, we had a test set that consisted of 300 random audios from our data set manually labeled by humans. We ran the classification model on that test set, achieving an F1 score of 0.97, a recall of 0.97, and a precision of 0.98, suggesting that the model worked well in our audio messages despite the model being optimized for the French language.

[15] https://github.com/ina-foss/inaSpeechSegmenter.

[16] We did not found any audio classified as a random noise/inactivity sample in our dataset.

Table 3.2. Distribution of Speech and Musical messages

Period	Class	Gender	%
Truck Drivers' Strike	Speech	Male	75.8%
		Female	15.8%
	Music	-	8.4 %
Election Campaign	Speech	Male	65.1%
		Female	18.0%
	Music	-	16.9%
Whole collected period	Speech	Male	65.6%
		Female	18.6%
	Music	-	15.8%

Table 3.2 shows the percentage of speech and music messages in the complete dataset. In all periods, we have a predominance of male speakers, reaching almost 76% in the Trucker Strike period while female speakers are between 16-18% in both periods. The music messages doubled during the Electoral Campaign period, mainly due to the many dissemination of "Electoral jingles", which is a common campaign method in Brazil to promote politicians during their campaigns.

3.4 General Characteristics

Here we outline the general characteristics of the dataset. Firstly, Figure 3.2 shows the number of audio messages shared daily during the period of collection. There is a peak of activity in May (in blue), which coincides with a national truck drivers' strike (between May 21st and June 2nd) that generated a lot of social mobilization in the country. We note a steady increase in the volume of audios shared by the end of the data collection (marked in orange), which coincides with the period of the general election campaign in Brazil (from August 16th to October 28th). On average, audios were shared 218 times a day, peaking at 1121 on the day before the voting day (October 27th). Considering the greater volume of audio messages shared in the two highlighted periods (blue and orange periods in the figure), we focus our analyses only on audio content shared during these periods.

Table 3.3 shows overall statistics about the data for the two selected periods as well as the whole collected period. It shows the total number of audio messages shared as well as the total numbers of users who shared at least one audio and groups where at least one audio was shared. Overall, we have more than eight thousand different users who shared almost 43 thousand audio files in 364 different groups. The table also

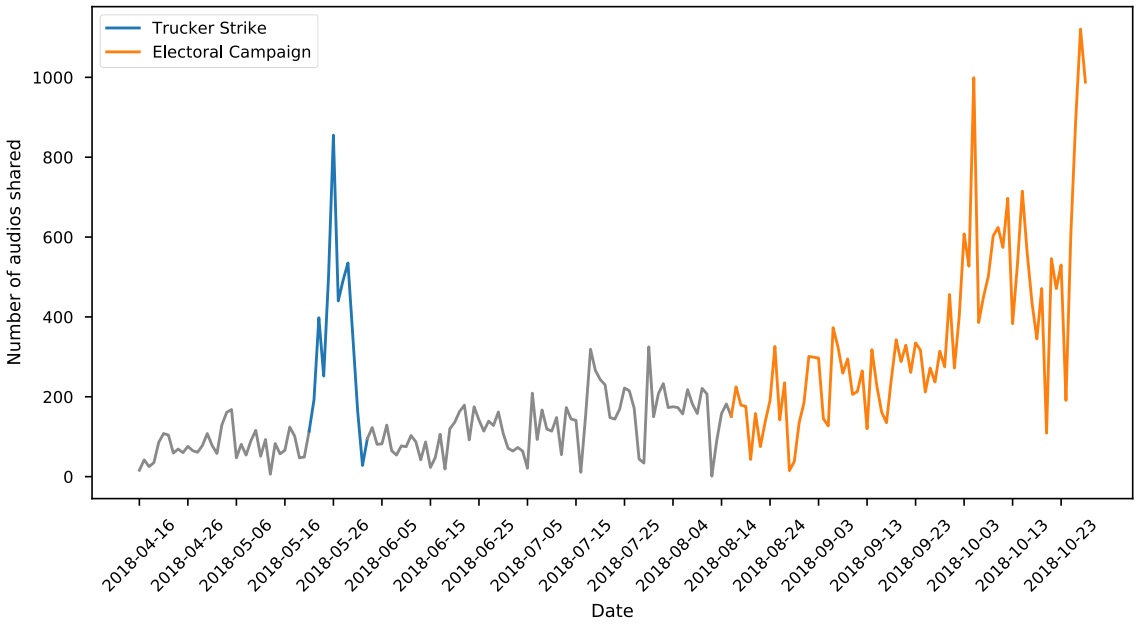


Figure 3.2. Number of audio messages shared in WhatsApp groups during monitored period.

Table 3.3. Dataset overview (* users and groups with at least one audio message).

Period	Type	Quantity
Trucker Drivers' Strike	# Groups*	117
	# Users*	1,134
	# Audio messages	5,780
	# Unique Audios	1,450
Election Campaign	# Groups*	330
	# Users*	6,002
	# Audio messages	28,593
	# Unique Audios	8,505
Whole collected period	# Groups*	364
	# Users*	8,056
	# Audio messages	42,869
	# Unique Audios	16,503

shows the total number of unique audio contents, which may indicate to potentially different audio files that convey the same content. We note that each audio content was shared 3-4 times on average, although, as we will see later, some audio contents were shared a larger number of times.

In general, roughly 32% and 21% of all users active in the monitored groups

during the electoral campaign and truck drivers' strike periods, respectively, shared at least one audio message. The fraction of monitored groups with at least one audio content also increased from 83%, during the truck drivers' strike to 90% during the election period. These numbers illustrate the increasing user participation in sharing audio content within WhatsApp groups.

3.4.1 Category Analysis

To get a better understanding of what was being discussed, we relied on 20 volunteers to manually annotate a sample of audios. Specifically, we randomly selected 100 audios from the truck drivers' strike period, 100 audios from the electoral period, and 100 audios from the top 500 audios most shared in our dataset, adding up 300 different audios. We asked the group of volunteers to categorize each audio into eight categories to get a glimpse of the information they conveyed. For each sampled audio, we required three annotations from three different volunteers. Finally, the volunteers were instructed to select all categories that fit their content. For comparison purposes, we adopted the same eight categories used in [Resende et al., 2019a], however, we listened to a great share of the audios to see whether additional categories were present. We found that no new categories stood out, thus we keep the original ones, which are:

- **Opinion**: a content expressing the speaker's opinion;
- **News**: information about an event, quoting or referencing a newspaper, magazine or news portal;
- **Politics**: information related to a candidate or party to publicize or praise some political subject;
- **Advertising**: commercials or ads related to a product, venue or service;
- **Satire**: Humorous content about current events or people;
- **Activism**: content encouraging or mentioning social movements, protests or other events
- **Inappropriate**: Hate speech, pornography;
- **Others**: content does not fit any other category;

The average Fleiss' kappa considering all categories is $\kappa = 0.49$, which indicates a moderate agreement. Considering each category individually, the one that had the

lowest agreement is *Inappropriate* with $\kappa = 0.18$ (therefore it was filtered out), while the category with the highest agreement is *Politics*, with $\kappa = 0.78$. Table 3.4 displays Fleiss' kappa for each category listed.

Table 3.4. Fleiss' kappa for each category

Categories	Fleiss' kappa (κ)
Opinion	0.50
News	0.37
Politics	0.75
Advertisement	0.40
Satire	0.61
Activism	0.42
Inappropriate	0.18
Others	0.66
mean	0.49

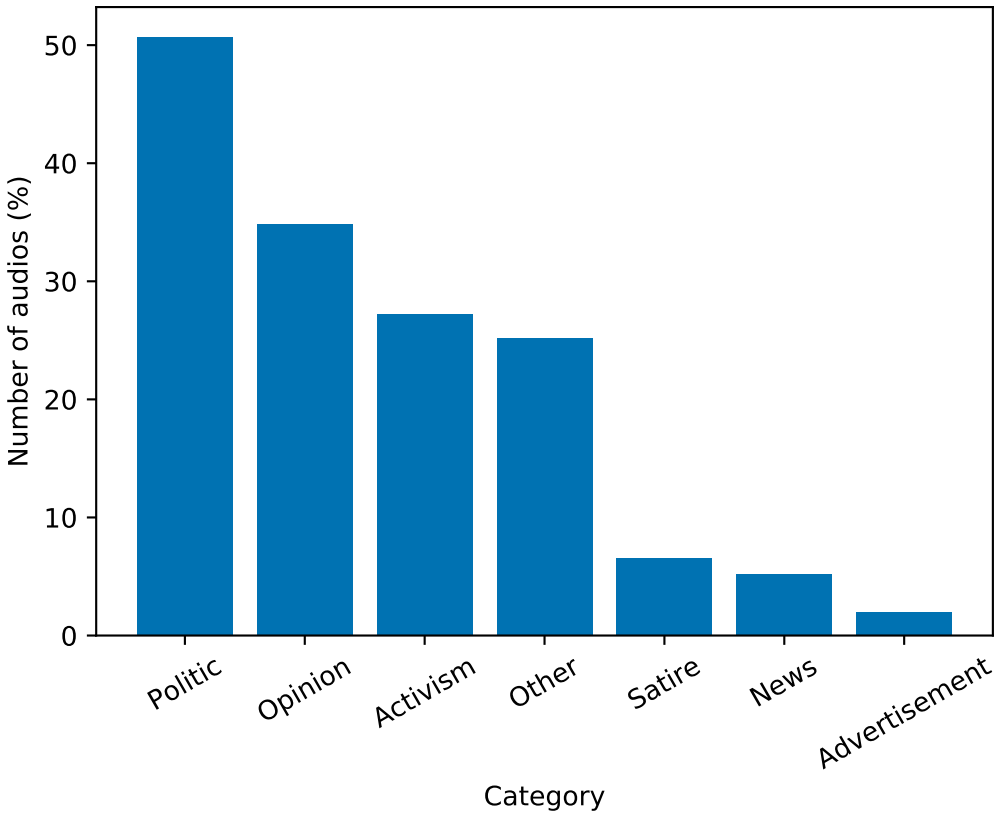


Figure 3.3. Distribution of audios across categories.

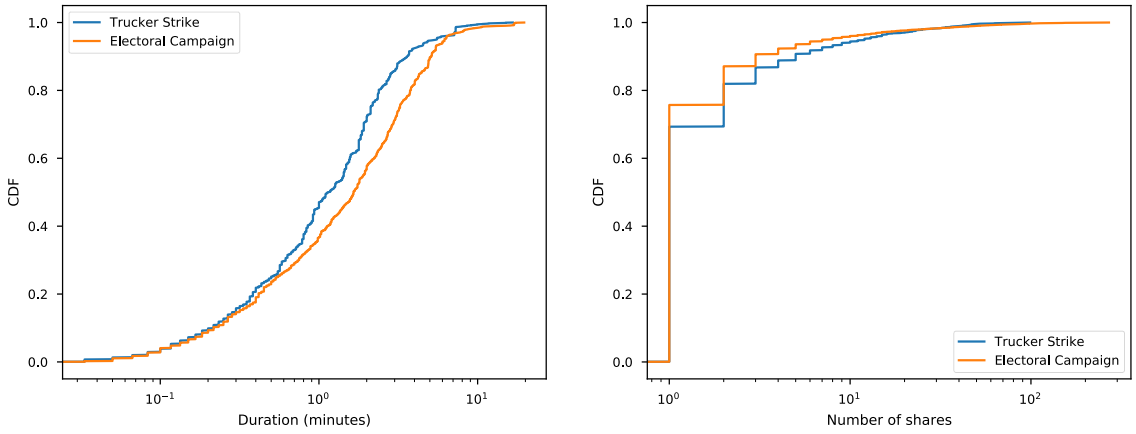
Figure 3.4. Wordcloud from the four most frequent categories (Translated from Portuguese)

Figure 3.3 shows how the 300 sampled audios are distributed across the eight categories. Note that the sum exceeds 100% as an audio may have been associated with more than one category at once. For instance, most audios labeled as opinions were also labeled as politics. The main categories of audios are *Politics*, followed by *Opinion* and *Activism*. Audios labeled as in *Others* category mostly relate to religious content, specific events, or unrelated chatter. *Satire*, *News*, and *Advertisement* categories appeared in less than 10% of audios. Compared to a similar categorization of image messages reported in Resende et al. [2019a], we observe a much larger presence of personal opinions and activism related content among the audio messages but less frequent use of this type of media to spread satirical content. These results illustrate important differences in how different media types are used to disseminate content in WhatsApp.

Figure 3.4 shows word clouds of the audio transcriptions for the most frequent categories in the annotation. Each word cloud represents the most frequent words in the audio (e.g., the more frequently a word occurs, the larger is its size). “Bolsonaro”, “Brazil”, “PT” (Workers Party), and several other words related to politics appear very frequently on both the Politics, Activism, and Opinion categories, indicating that this subject was frequently discussed. Specifically for the Activism category, we see

that “Trucker drivers" is well represented due to the Trucker’s Strike between May and June. The “Other" category contains religious references, such as “God" and “Life" as well as words related to casual chatter.

3.4.2 Audio Duration and Number of Shares



(a) CDF of audio duration

(b) CDF of number of audio shares

Figure 3.5. CDF of Audio duration and number of shares

Figure 3.5 shows, for both analyzed periods, the cumulative distributions (CDF) of the durations. The average duration is around 2 minutes for both periods, though audios shared during the election period tend to be somewhat longer: around 20% of the audios shared during that period have more than 3.5 minutes (versus 2.5 minutes during the strike). Only 139 unique audios are longer than 20 minutes. Figure 3.5 also shows the CDFs of the total number of times each audio message (i.e., all audios grouped as similar content) was shared in all monitored groups in both periods analyzed. As shown, some audios have a vast reach: for instance, 10% of the audios were shared more than ten times during the election campaign, and the audio that appeared most times had 270 shares. We computed the Pearson and Spearman correlation and coefficient between the number of shares and the duration of the audio but found no noticeable correlation.

3.5 Summary

In this chapter, we presented the methodology, composed of six main steps: the WhatsApp dataset collection which used the *WebWhatsapp-Wrapper*[17], pre-processing using textitpyDub[18], grouping audios with similar content with *Chromaprint*[19], audio transcription using Google Cloud's Speech-to-Text API[20], misinformation detection and finally, audio type categorization between speech and music using the inaSpeechSegmenter[21]. We finished by presenting a brief characterization of the data collected and processed. The next chapter presents an analysis of the data, contrasting properties of the audios with previously checked misinformation from the rest of them.

[17] https://github.com/mukulhase/WebWhatsapp-Wrapper.
[18] https://github.com/jiaaro/pydub.
[19] https://acoustid.org/chromaprint.
[20] https://cloud.google.com/speech-to-text/.
[21] https://github.com/ina-foss/inaSpeechSegmenter.

Chapter 4

Content and Propagation Dynamics Characterization

In this chapter, we go over our main findings regarding the content of audios shared in WhatsApp. We start by looking into specific details of the content shared, such as the topics discussed, and which attributes they have in them. We then present a qualitative analysis with a group of people, to identify specific properties of the audios. Finally, we analyze the propagation dynamics, discussing how they spread across the network. In all analysis we contrast the differences between audios with previously checked misinformation and unchecked content.

4.1 Audio Content Analysis

In this section, we analyze the content of the audios shared in the WhatsApp groups. We here focus on the content of the audio messages. To that end, we focus on the transcriptions. We perform a topic analysis uncovering the main topics of discussion conveyed in the shared audios, and then we look into some psychological linguistic features extracted from the transcriptions.

4.1.1 Topic Analysis

We further characterized the audio transcriptions in terms of the distribution of the topics they talk about. We used the model Latent Dirichlet Allocation (LDA) [Blei et al., 2003], a generative statistical model to automatically infer the topics in a collection of documents D, in our case the audio transcription, to infer the topic distribution of the audio messages. LDA receives the audio transcriptions of all audio messages

and the desired number of topics k, and it computes the topic distribution, which can be interpreted as k clusters of words. For each audio transcription a, we can infer which topics are discussed. To find the optimal number of topics, metrics such as topic coherence [Röder et al., 2015] can be calculated by a range of possible topics. The number of topics with the highest topic coherence is chosen.

As a first step, we applied the pre-processing phase to the transcriptions by removing punctuation marks and stop words (common words, such as "the", "a" and "an"), lowercasing all the words, and by applying lemmatization (removing inflectional endings and returning the base form of the word, e.g., cats becomes cat, caresses become caress) to make inflected words comparable to each other. These steps are important as studies suggests that these pre-processing procedures improve the results of supervised and unsupervised text-analysis techniques, including topic modeling [Denny and Spirling, 2017]. To apply this pre-processing step, we used SpaCy[1], a natural language processing library in Python that can be applied to the Portuguese language.

After all the transcriptions are pre-processed, we pass them as input to the LDA model. We used the implementation provided by gensim[2], a topic modeling library for Python that has the LDA algorithm implemented as well as topic coherence metrics. The model returns the words associated to the k topics learned by the model, and with those words, we can get a better understanding of what is discussed in each topic.

The LDA requires a set of hyperparameters, specifically the number of topics that we will infer from the transcriptions. To find out the best number of topics, we ran the algorithm varying the number of topics k from 2 to 20 and assessed the quality of the results. For the assessment, we measured the topic coherence c_v of the results. The topic coherence metric captures whether different topics have actual few words in common. For that objective, it uses the count of co-occurrences of words, pointwise mutual information (NPMI), and cosine similarity [Röder et al., 2015].

Figure 4.1 shows the coherence score for each number of topics (k). We found the best topic coherence at $k = 8$ topics. Table 4.1 presents the most representative words for each topic. Note that topics 1, 3, and 5 are closely related to politics since they are characterized by words such as "Campaign", "Brazil", "Mayor", "Politician", "PT" and so on. Topic 4 is closely related to the Trucker Strike event, identified by the words "Trucks", "Truck Drivers" and "Military Intervention" (a topic largely discussed during the trucker strike movement). Topic 6 contains mostly words related to religion, suggesting that many audios were recordings of members of Christian denominations members. Finally, topics 2, 7, and 8 are more loosely connected and encompass more

[1] https://spacy.io/.
[2] https://radimrehurek.com/gensim/.

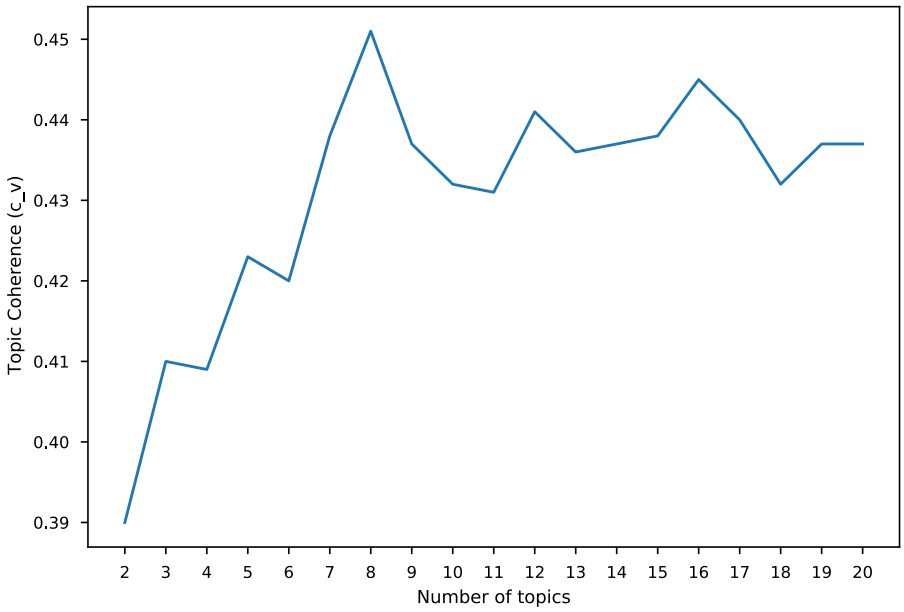


Figure 4.1. LDA Topic Coherence

general narratives.

Table 4.1. Most representative words for each topic inferred by LDA method

Topic	Most representative words
1	Brazil, Country, Person, Brazilian, Politician, Year, PT, Family, Govern, Defend
2	Expensive, Talk, Stay, See, Understand, Marry, End, Impose, Woman, Nobody
3	Federal, Public, Congressperson, Million, Lula, Paulo, Money, Year, Candidate, Politician
4	Military, Brazil, Stop, Trucker Driver, Army, World, Brazilian, Military intervention
5	Bolsonaro, Vote, Brazil, Haddad, PT, President, Jair, Election
6	God, Lord, Jesus, Life, Word, Day, Love, Heart, Father, Name
7	Guys, People, Talk, Understand, Stay, Do, Happen, Find
8	Day, Hour, Guys, City, Car, Night, Today, Come, Friend

In order to analyze the distribution of topics across different audio transcriptions, we first assigned to each transcription the most prevalent topic according to LDA results (i.e., the topic with highest probability associated with the transcription). Figure 4.2 presents the distributions of topics across different transcriptions, separately con-

sidering audios with misinformation and audios with unchecked content. Note that 52% of audios with misinformation are characterized as containing content related to topics 4 and 5, which are the most politically oriented topics and include words such as “Military”, “Trucker Driver”, and “Bolsonaro”, and “Election”. Topic 7 is the third most predominant topic among audios containing misinformation: 18% of them are characterized by this topic which covers words such as “Guys”, “Understand”, and “Find”. Due to it being common starting words, many phrases start with “Guys, you need to understand...”, or some other variation. Unchecked audios are more equally distributed across all topics. The topic that holds the highest audios with unchecked content is Topic 2, which is characterized by words like “Expensive”, “Understand” and “Talk”, with almost 21% of audios falling into this category.

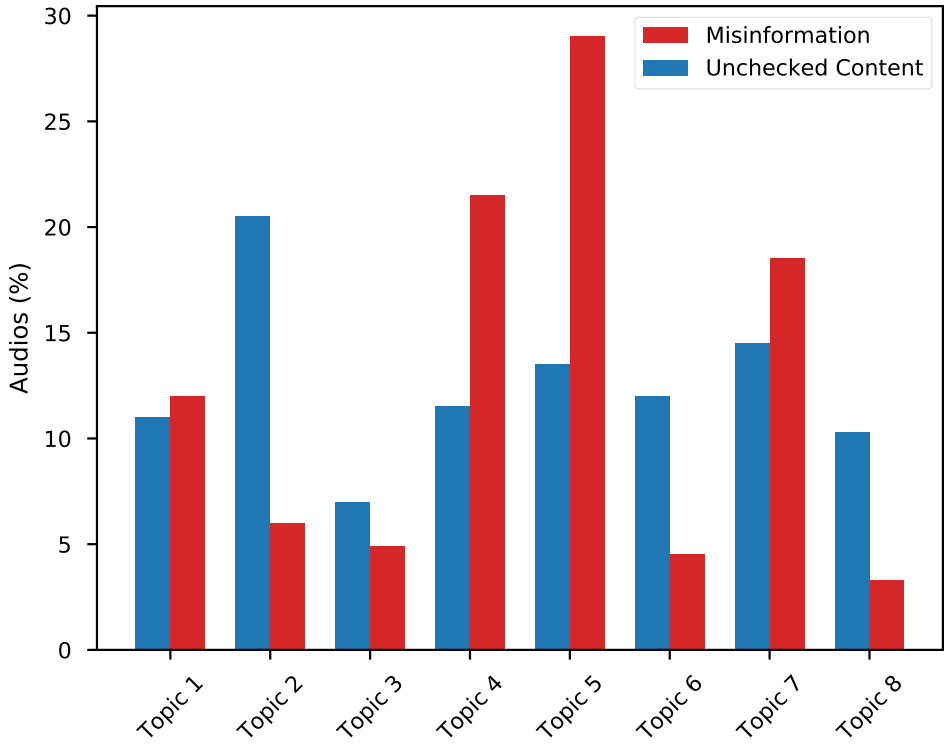


Figure 4.2. LDA Topic Distribution

4.1.2 Psychological Linguistic Features

To extract the distribution of psychological linguistic features from the audio transcriptions, we used the the 2015 Linguistic Inquiry and Word Count (LIWC) [Pennebaker et al., 2015] lexicon. LIWC is a dictionary containing word categories (or *attributes*)

associated with emotions, thinking styles, social concerns, and even parts of speech. It contains a text analysis module that takes as input written or transcribed verbal texts, compares each input word with the pre-defined attributes and produces as output the percentage of all input words that match each of the LIWC attributes. We use the Portuguese dictionary[3], which has in total 41 word categories, or attributes. Examples include attribute *negemo* (negative emotion), characterized by words like "hate" and "ugly", as well as attribute *future*, characterized by words "will" and "soon".

We passed each audio transcription to LIWC, which in turn returns a value indicating how present each of the attributes provided is in the text. An audio transcription with hate speech, for example, would have the attribute *negemo* (negative emotions) and *anger* with a high value and attributes such as *posemo* (positive emotions) with a value close to zero.

With the attributes calculated for every single audio transcription, we compared the distributions of each of these attributes against audios that were marked as containing misinformation versus audios with unchecked content. This comparison aimed to identify which attributes were significantly different in these two types of audio messages. We applied the Kolmogorov-Smirnov to these distributions and selected the attributes that had a p-value < 0.05.

Aiming at contrasting the most common LIWC attributes on audio transcriptions classified as misinformation and transcriptions containing previously unchecked content, we computed, for each LIWC attribute that was marked as being significantly different in the two types of messages, the ratio of the difference between the values of the attribute in audios with misinformation and in audios with unchecked content to the value of the attribute in audios with unchecked content. Figure 4.3 shows the relative differences of these attributes. A positive difference means that messages with misinformation had more of that attribute than those unchecked and vice-versa.

From Figure 4.3, we note that audios with misinformation have a higher word count (WC). Furthermore, messages with misinformation tend to be more related to work, with work-related words such as jobs and employment, have more negative emotions (e.g., hate, ugly, worried), use words from the third person singular, such as she and he, carry phrases in the future tense and have words related to insights, such as "think" and "know". Moreover, audios with misinformation also tend to use words such as "you" or "your" (e.g., "it is your problem") and use words related to causation, such as "because" and "to that effect". We display a few examples from the most relevant attributes in Table 4.2.

[3]Provided by http://143.107.183.175:21380/portlex/index.php/pt/projetos/liwc.

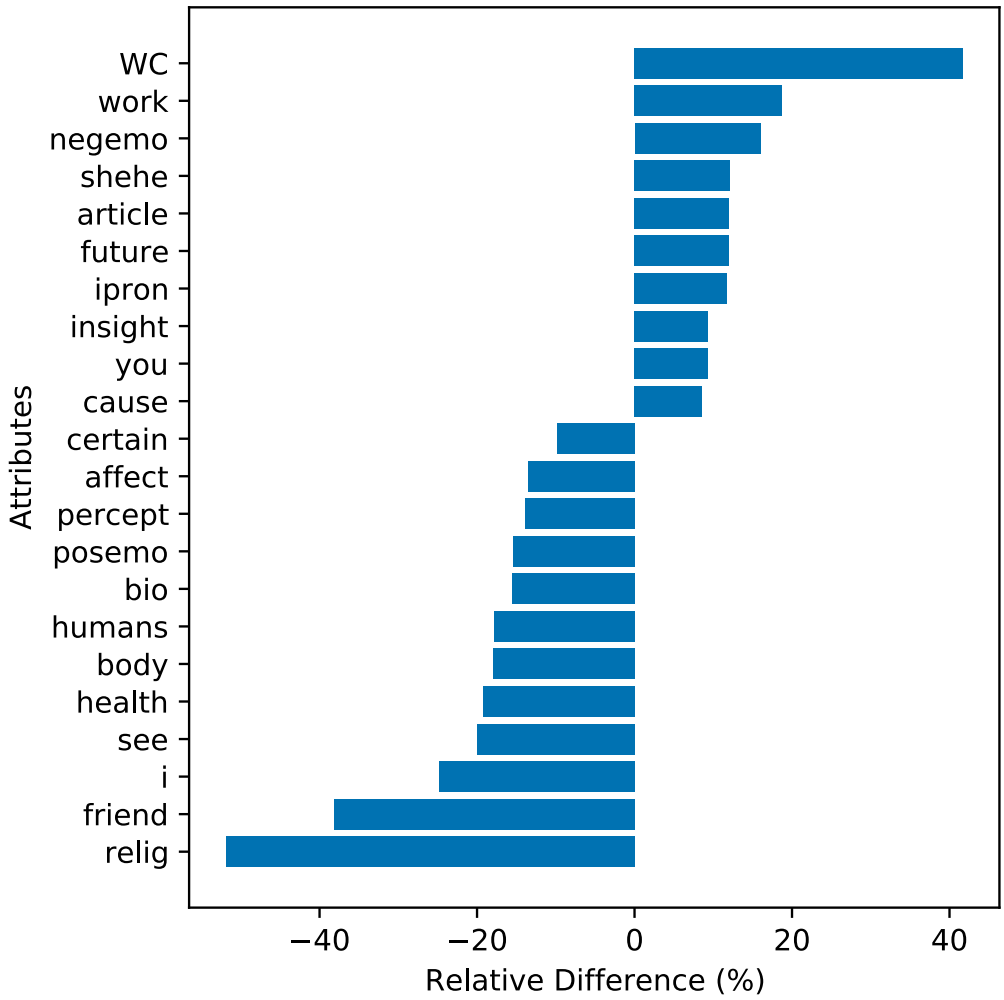


Figure 4.3. Relative difference between audio messages with misinformation vs. with unchecked content.

On the other hand, religious content are more present in audios with unchecked content, as well as friendship-related words (e.g., “friends”, “buddy”), health and biological words (e.g., “hospital”, “flu”, “body”). Moreover, in contrast to audios with misinformation, audios with unchecked content tend to have a higher predominance of positive emotions such as “nice”, “sweet” and “love” associated with attribute *posemo*. A complete description of every attribute can be seen on Tables 4.3 and 4.4.

When comparing these results with those obtained for textual messages by Resende et al. [2019a], we notice that in both audio and textual messages, there is the predominance of the *insight* attribute on misinformation, characterized by words such as “attention”, “warning” and “listen”. However, Resende et al. [2019a] pointed out that words such as “we” and “they” appear more frequently in misinformation, which

Table 4.2. Examples of transcriptions (Translated from Portuguese)

Attribute	Example
work	[...] Lindbergh Farias receives an absurd amount of money each month and **is roaming in Curitiba instead of working at the Congress** [...] [...] I saw bandits being victimized and the **working citizen** arrested, held hostage to violence, I saw the schools violate the innocence of our children [...]
negemo (negative emotion)	[...] the news that they are spreading and trying to connect it to the CPMF is a lie; it is just another lie **from these bastards** [...] [...] please share this as much as you can to arrest, **this scoundrel** is promising **to kill** 30 to 40 children [...]
future	[...] the great nations of the world and the most advanced in the world **will never accept** electronic voting machines [...]

Table 4.3. Positive Relative Difference LIWC Attributes

Attribute	Description	Keywords
WC	Word Count	-
work	Work related words	job, company, employment
negemo	Negative Emotions	hurt, ugly, nasty, hate, worried
shehe	3rd pers singular	she he
article	Article	the, a, an
future	Future tense	may, will, soon
ipron	Impersonal pronouns	One, They, You, It
insight	Cognitive Process of Insight	think, know
you	2nd person	you, your
cause	Causation	because, effect

is not the case here, with "you" being more frequent in audios containing this type of content. Textual messages with misinformation also had a high presence of the *sexual* attribute, corresponding to words such as "nudism" and "sex", however we found no significant presence of this attribute in audio messages with misinformation. Lastly, textual messages also tend to have more words associated with the present, whereas audios are more often associated with sentences in the future tense.

This attribute analysis shows that audio messages and textual messages have similarities but also have unique characteristics, suggesting that the approach or method of sharing misinformation varies depending on the type of media used. Even though both media types use the *insight* attribute, textual messages are more aimed at aggregating the community towards the same go and often refer to third-parties as collectives as well by using "we" and "they". They often talk in the present tense. Audio messages

Table 4.4. Negative Relative Difference LIWC Attributes

Attribute	Description	Keywords
certain	Certainty	Always, Never
affect	Affective Process	Happy, Cried
percept	Perceptual Process	look, heard, feeling
posemo	Positive emotions	love, nice, sweet
bio	Biological Process	eat, sleep, blood
humans	Humans	Adult, Baby, Boy
body	Body	Cheek, hands, spit
health	Health	Clinic, Flu, Pill
see	See	View, saw, seen
i	1st pers singular	I, me, mine
friend	Friends	Buddy, Friend, Neighboor
relig	Religion	Altar, Church, Mosque

are more target to the listener itself by using "you" more often. Audio messages also focus on the future tense.

4.2 Qualitative Analysis

Aiming at delving deeper into the content of the audios, we conducted a qualitative analysis in a sample of 100 audios based on two phases: an interview and a survey. The purpose of the interview was to gather the perceptions of selected volunteers on the audios' content and the potential feelings they could infer from the speaker's voice. Using the insights from the interview phase, we developed an online survey with a pre-defined set of questions to reach a broader public [Fraser and Gondim, 2004; DiCicco-Bloom and Crabtree, 2006]. We describe these two phases in the following sections.

4.2.1 Interview

We interviewed three volunteers separately, with each interview consisting of a one-hour session via Skype, using a semi-structured format, with a defined list of questions that were to be followed by the volunteers. Each interview can be divided into two phases. The first phase consists of questions aimed at learning more about the volunteer, their participation in WhatsApp groups, and their perception of audio contents in general. During the second phase, the volunteers were asked to listen to four different, randomly selected audio files, two with misinformation and two unchecked, followed

by questions regarding their perception of each audio. The volunteers did not know if the audio content contained misinformation or unchecked content. Each volunteer received a consent form to allow the use of their responses in this study in an anonymous format. The questions that compose each interview were created based on qualitative analysis about fake news and misinformation from other authors, such as Wagner and Boczkowski [2019]; Roozenbeek and Van Der Linden [2019]; Zhou and Zafarani [2018], and are displayed in Table 4.5.

Table 4.5. Questions in the initial interviews

Block of questions	Questions
Profile	- Name - Age - Educational Background
General Questions	- How many audios you usually send or share in WhatsApp groups - What is the largest WhatsApp group that you are a member of? - If an audio is spoken by a known person (e.g. celebrity), would it have more credibility? - If an audio is spoken by a friend or family, would it have more credibility?
Specific to each audio	- What is your level of knowledge about the audio subject? - What emotion do you have when listening to the audio? - Would you share this audio? Why? - What is your level of knowledge about the topic of the audio?
Audio Content Details	- Did you notice anything peculiar about the audio? Do you think the person speaking has knowledge about what he/she is talking? - Was there any word or part of the audio that caught your attention? - Was there any background noise? Do you think the audio was edited?

The three volunteers were in the age group of 25-35 years old and had majored in computer science or system analysis. Two of the volunteers had previous studies published involving some misinformation studies while one of them had not. The three of them were in groups with more than 50 members, some of which had even more than 100 members (usually groups with university students to discuss general matters such

as possible rides to/from the campus). They also pointed out that they did not often send audio messages but did receive them on a daily basis. However, they reported that they did not always listen to the audios due to time restrictions. Finally, all three volunteers said that in their opinion, audios recorded by a public personality, such as a celebrity, do not necessarily have more credibility. However, they would pay more attention to audios shared by friends or people they trust.

Regarding the volunteers' perception of the listened audios, we analyze their answers separately for audios with misinformation and audios with unchecked content. Misinformation audios were spotted easily as potential sources of misinformation, possibly due to the volunteers' close relation to misinformation studies, which was not a problem as the main focus of this phase was to raise the main characteristics that were evoked. In some cases, they also reported finding a certain tone of artificiality in the speakers' tone. In the following, we present a list of the remarks and general insights from the volunteers:

- Many audios containing misinformation created uncertainty on the listeners. They were unsure about the veracity of the fact of what was being said.
- Often the speaker of audios with misinformation tried to create a link with someone important (e.g., they knew the owner of a oil company, were related to a famous newscaster).
- Audios with misinformation tried to back their claims with sources but they were not considered reliable by the volunteers.
- Some audios with misinformation were cited as sounding like conspiracy theories, often citing some major event that was happening or was about to happen.
- Audios with misinformation tried to engage more with the listener, often trying to create the illusion of familiarity and intimacy with the listener, calling them friends or family.
- The listeners mentioned feeling anger and disgust when listening to audios that had misinformation.

Overall, we noticed several peculiarities in audios with misinformation, such as the feeling of uneasiness or ways to try to engage with the public, which could be frequent in audios in this category. With these peculiarities mapped out, we moved into the second part of the analysis to identify whether these characteristics were frequent.

4.2.2 Online survey

Based on the insights collected in the interview with volunteers, we set up an online form with a set of questions to gather more information on differences between audios with misinformation and unchecked content, as perceived by a larger group of listeners (here referred to as volunteers). We also wanted to check whether the previous remarks collected were also noticed by this larger group and a broad set of audios. The online form is composed of two sets of questions, presented in Table 4.6. One set contains questions related to demographic data about the annotator and questions about the frequency of usage of audios in WhatsApp groups, and the other contains questions related to impressions they had after listening to a given audio message.

For the survey experiment, we selected a random sample of 100 audios, 50 with unchecked content and 50 with misinformation content, and asked people to answer our online form. We publicize the survey among friends and the members of the Social Computing Laboratory from Universidade Federal de Minas Gerais (UFMG). We left it open until exactly three different people evaluated each of the 100 audios. Five different audios with less than three annotations were randomly assigned to each volunteer upon entering the survey webpage. If the volunteer desired, he had the option to display more audios than the initial ones assigned.

In total 25 volunteers participated in the online survey. We describe the observations from the survey, starting with the first set of questions (question 1-5). Regarding the ages of volunteers, on average, they were 27 years old (question 1). Moreover, 16 volunteers identified as male and 9 identified as female (question 2). The sizes of the largest group they participated in (question 3) varied from 25 to 256 (the maximum size of WhatsApp groups) members, with average 88, which coincides with previous observations that WhatsApp groups tend to connect large group of people [Resende et al., 2019a; Caetano et al., 2019]. For question 3, the average size of the largest group that they participated in was 88 people. The smallest group had 25 people and the largest 256, which corroborates with the statement that WhatsApp connects large groups of people, much like social media. As for the number of audios received on WhatsApp daily (question 4), three volunteers said they receive no audios, sixteen received 1 to 5 audios a day, five received 6 to 10 audios, and one volunteer reported receiving more than ten audios a day. As for the number of audios shared on WhatsApp daily (question 5), twelve volunteers said they sent no audios, and thirteen said they send 1 to 5 audios daily. Thus, in general, our volunteers tend to receive more than send audios on WhatsApp.

For the second segment of the survey (questions 6-15), each answer was tied to a

Table 4.6. Questions in the online survey

	Questions	**Possible Answers**
1	What is your age?	Open question
2	To which gender identity do you most identify?	Male, Female, Other, Prefer not to disclose
3	What is the maximum number of members in the WhatsApp groups you are a part of?	Open question
4	What is the approximate quantity of audios you listen to or receive every day in WhatsApp?	0, 1 to 5, 6 to 10, 10+
5	What is the approximate quantity of audios that you share every day on WhatsApp?	0, 1 to 5, 6 to 10, 10+
6	Which emotion did you feel when listening to the audio?	Sadness, Surprise, Fear, Trust, Joy, Anticipation, Anger, Disgust, Other
7	Do you think this audio contains false information?	Yes, No
8	The audio contains some data or source that tries to support the content?	Yes, No
9	If the previous question is true, does this source increases the credibility of the information?	Yes, No
10	Would you share the audio with any of your contacts? If so, why?	Yes, No, Open question
11	How natural is the person speaking the audio?	Likert Scale: 0 (Very Artificial) to 4 (Very Natural)
12	How excited is the person speaking the audio?	Likert Scale: 0 (Very Sad) to 4 (Very Excited)
13	How friendly is the person speaking the audio?	Likert Scale: 0 (Very Hostile) to 4 (Very Friendly)
14	Does the audio have any calls to actions, for instance sharing the content?	Yes, No
15	Was there anything else that caught your attention (Optional)	Open Question

specific audio file, and we separate our analysis based on whether the audio contained misinformation or unchecked content. Figure 4.4 shows the distributions of emotions felt by the volunteers when listening to the audio files (question 6). For each audio, we selected the emotions felt from all the volunteers. Based on the answers, we can see that the participants felt more negative emotions when listening to audio messages

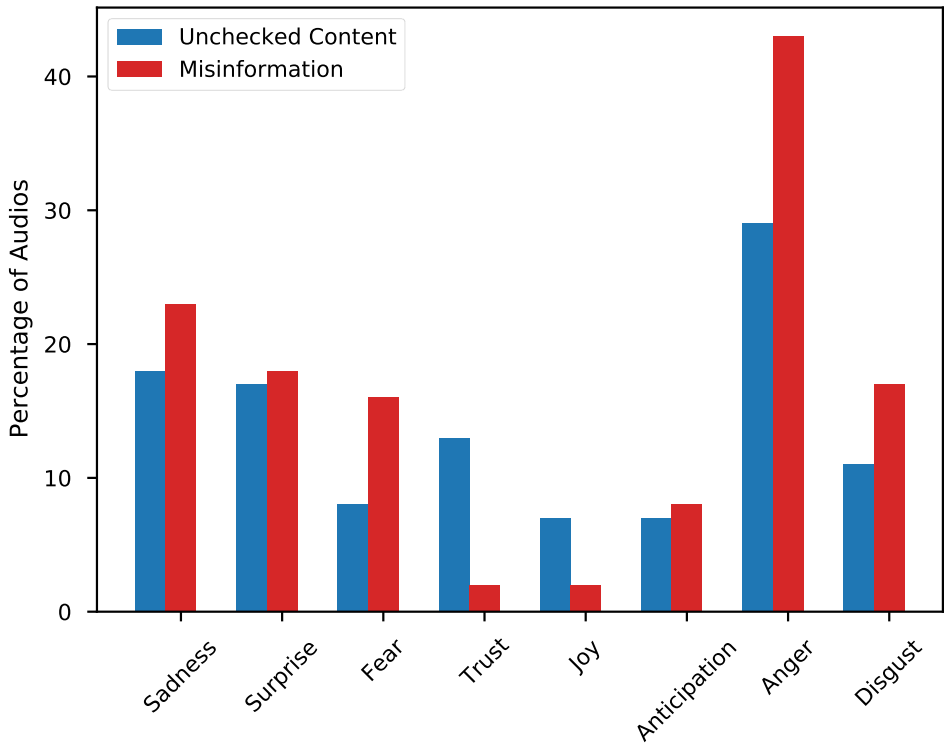


Figure 4.4. Distribution of emotions felt by volunteers when listening to audios in different categories (misinformation or unchecked content)

with misinformation. This might be due to the higher presence of negative emotion-related words, which we found in the analysis done in Section 4.1.2, regarding the psychological linguistic features. Sadness, surprise, fear, disgust, and especially anger were most felt while listening to audios with misinformation, whereas trust and joy were most reported when listening to audios with unchecked content.

As to question 7, when presented audios with misinformation, our volunteers spotted them 76% of the time. When presented with unchecked audios, 43% thought that the audio contained misinformation. When asked whether the audio had some form of data or source to back the information (question 8), 58% responded yes when presented an audio with misinformation, and only 17% responded yes when presented an audio with unchecked content. However, when asked whether the source provided increased the credibility of the audio (question 9), only 24% of the volunteers said it does indeed increase the credibility. This reaffirms some points raised by the volunteers in Section 4.2.1: many audios with misinformation try to back their history with some study or data, but they are often not reliable enough. This also links back to the *insight* attribute found on audios with misinformation in Section 4.1.2, where insight words

(e.g., think, consider, know) are often used when trying to create this storyline. Most of the volunteers said they would not share any of the audios (question 10): for audios with misinformation, only 9% of the volunteers mentioned that they would share them with friends or family just to comment on them, whereas for audios with unchecked content, this fraction drops to 5%.

We found a significant difference in the answers for audios with misinformation and unchecked content regarding the *friendliness*, and *naturality* of the speakers from both types of audio, but we found no significant difference for *excitment* (questions 11-13). The friendliness score given to audios with misinformation was, on average 1.78 and 2.34 for audios with unchecked content. These two score are statistically different according to a t-test with p-value ≤ 0.05. Thus, in general, speakers in audios with misinformation are perceived as less friendly than audio speakers with unchecked content. As to the naturality of the speaker, the gap is even larger. Misinformation audios had an average score of 1.65 while unchecked audios had 2.56 (statistically different according to a t-test with p-value ≤ 0.05), suggesting that speakers in audios with misinformation tend to more often pass the impression of some artificial tone.

We also found out that audio with misinformation is often accompanied by some call on action (question 14). That is, volunteers reported observing some form of instruction to be executed by the listener (e.g., share the audio in more groups) in audios with misinformation in 72% of the cases. For audios with unchecked content, this fraction falls to only 32%. Regarding the last open question (question 15), the volunteers noted that some audios with misinformation were very hostile. They also had the impression that the speakers tried to impersonate someone that would be trustworthy for that particular information (e.g., a nurse or someone from the military), or even say that the speaker had "privileged information" but cannot disclose the source. For unchecked content, the volunteers only pointed out how extensive some of the audios were.

In sum, the survey results suggest the following key observations: Audios with misinformation tend to make the listeners feel more negative emotions, such as sadness, fear, anger, and disgust. Audios with misinformation also were cited as having some source to try to support their claims, but these sources were often seen as unreliable and, in many cases, did not make the information more believable. The speaker's tone of audios with misinformation was considered less *friendly* and less *natural* than the audios with unchecked content. Finally, the volunteers also noted that audios with misinformation were also more accompanied by some form of instruction to be executed by the listener, such as sharing the audio to other groups, which was not the case for audios with unchecked content.

4.3 Propagation Dynamics

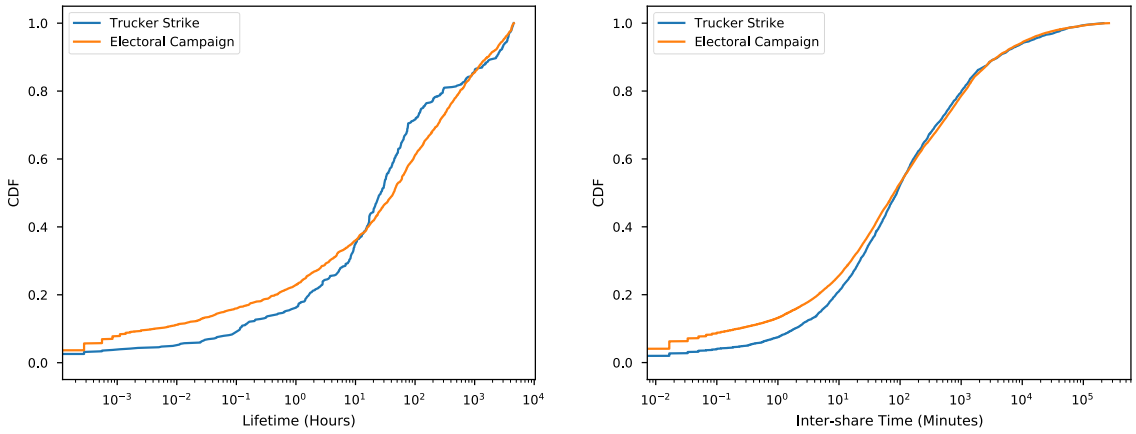
In this section, we look into the propagation dynamics of audio messages, looking into metrics such as lifetime and inter-share time. The former is the time interval between the first and the last times a particular audio content was shared in any monitored group, $t_n - t_1$, where n represents the number of times the audio was shared in any group, whereas the latter is the time interval between consecutive shares of the same content (regardless of the group in which it was shared), $t_2-t_1, t_3-t_2, t_4-t_3, ..., t_n-t_{n-1}$. We also look into how many groups each audio message reaches and how many unique users share the same audio.

In the following, we first analyze the propagation dynamics for the trucker strike and electoral campaign to understand if these events have unique characteristics as well as to contrast the propagation dynamics of audio messages versus textual and image content, which was explored by Resende et al. [2019a,b]. We then switch over to analyzing the whole collected period and go over the differences in the propagation of audio messages with misinformation and unchecked content. Finally, we analyze the differences between audio messages containing speech versus music and gender differences.

4.3.1 Trucker strike and Electoral campaign

Figure 4.5 shows the distributions of lifetimes and inter-share times during the trucker strike and the electoral campaign, separately to have an idea of similarities and differences in propagation dynamics during the two periods. As the figure shows, both distributions are quite similar in both periods.

As shown in figure 4.5(a), 50% of the audios stopped being re-shared after only one day since their first appearance. Moreover, according to Figure 4.5(b), roughly 60% of the audios are re-shared within 3 hours, and 20% are re-shared within 6 minutes, in both periods. These numbers are significantly different from those previously reported for textual messages Resende et al. [2019b]: audio messages tend to spread more slowly (longer inter-share times) but also remain for shorter periods in the system (shorter lifetimes). Such differences may reflect the greater effort required to listen to an audio message (compared to reading a text). In any case, it is interesting to note that a fraction of the audios remained in the system for quite some time: the lifetimes exceed ten days for roughly 20% of the audios.



(a) Lifetimes (b) Inter-share time

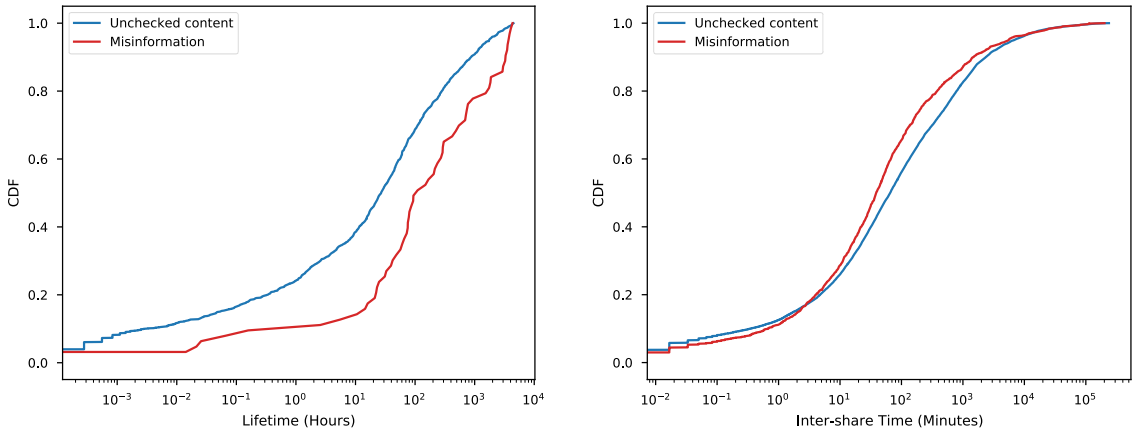
Figure 4.5. Distributions of lifetimes and inter-share times of audio messages in the trucker strike and electoral campaign

4.3.2 Misinformation versus Unchecked Content

Figure 4.6 shows the distributions of lifetimes and inter-share times for audios with misinformation and unchecked content, now considering audios shared during both periods. As shown in Figure 4.6(a), 75% of audios with unchecked content tend last at most seven days in the system, whereas the same fraction of audios with misinformation last up to 31 days, an increase of 3 weeks more than the unchecked content.

These numbers represent a significant increase compared to results previously obtained for image content from Resende et al. [2019a,b]. In these previous studies, the authors found that roughly 70% of images with misinformation tend to last in the system during about the same time as the unchecked images (100 hours). Textual messages on the other hand, had a more similar behavior to the audio content in which the misinformation content last longer in the network. Resende et al. [2019a] found that 50% of textual messages with misinformation last up to 10 days in the system. Here, we observe that the same fraction of audio content with misinformation last for up to 6 days.

Figure 4.6(b) shows distribution of inter-share times for both audios with misinformation and unchecked content. In this context, we see that the difference between these two distributions is more subtle when compared to lifetime. Roughly speaking, around half of the audios with misinformation are re-shared with 40 minutes whereas the same fraction of audios with unchecked content are re-shared within 65 minutes. Thus, audios with misinformation tend to spread somewhat more quickly than unchecked content. A similar behavior was detected in image and textual content



(a) Lifetime

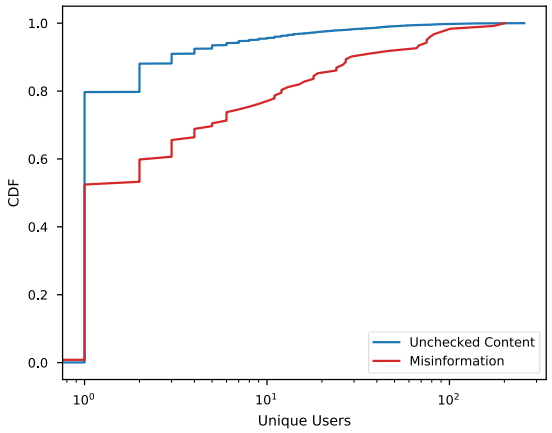
(b) Inter-share Time

Figure 4.6. Distributions of lifetimes and inter-share times of audio messages in audios with misinformation and unchecked content

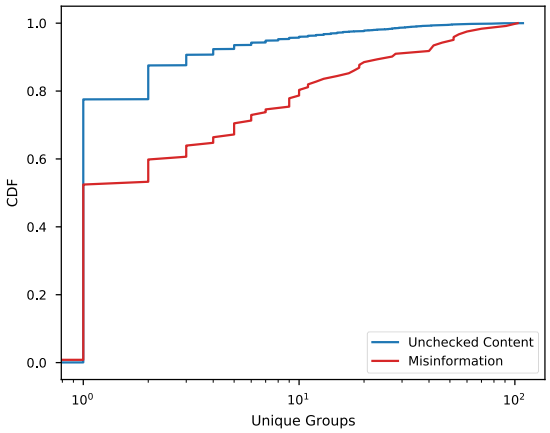
by Resende et al. [2019a,b]. However, image content with misinformation spread a lot faster than the audio content: according to Resende et al. [2019a], around 80% of the images with misinformation are re-shared within 100 minutes, but we found that 65% of the audios with misinformation are re-shared with the same time interval.

We now turn to the analysis of the reach of audio messages, in terms of users and groups, contrasting audios with misinformation and audios with unchecked content. Figure 4.7(a) shows the distributions of the number of users (unique users) who shared a specific audio message. Around 80% of messages with misinformation are shared at least by 12 different users, while 80% of unchecked audio are shared by at most two people. Figure 4.7(b) shows the distributions of the numbers of groups each message was shared in. Here, we have that 90% of audios with misinformation are shared to at least 27 different groups, while the same fraction of audios with unchecked content appears only in three groups. Figure 4.7 shows the distribution of the number of shares a audio message had. 80% of audios with misinformation were shared at most 13 times, while for the same fraction, unchecked content had a maximum share count of two. These numbers show the "viralization" properties and potential that audios with misinformation have over general, unchecked audio. This can be explained by factors such as:

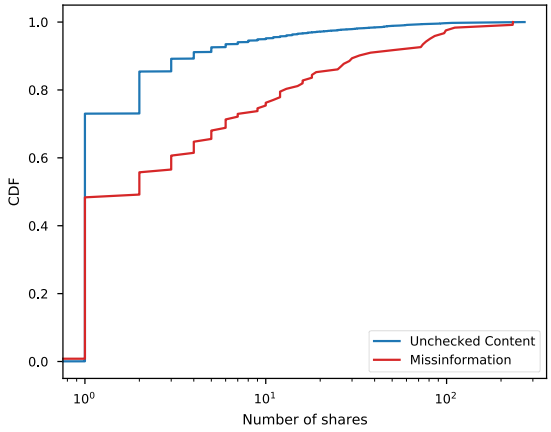
1. Audios with misinformation tend to target topics that are incredibly relevant to the current political scenario that they appear in, such as political candidate discussion at the electoral period, or involving major opinions toward strikes as seen in the topic analysis as seen in Section 4.1.1;



(a) Number of sharing users per messages



(b) Number of groups per messages



(c) Number of times shared per message

Figure 4.7. Distribution of Number of Groups per Message, Users per Message and Number of Times shared per message for Misinformation versus Unchecked content

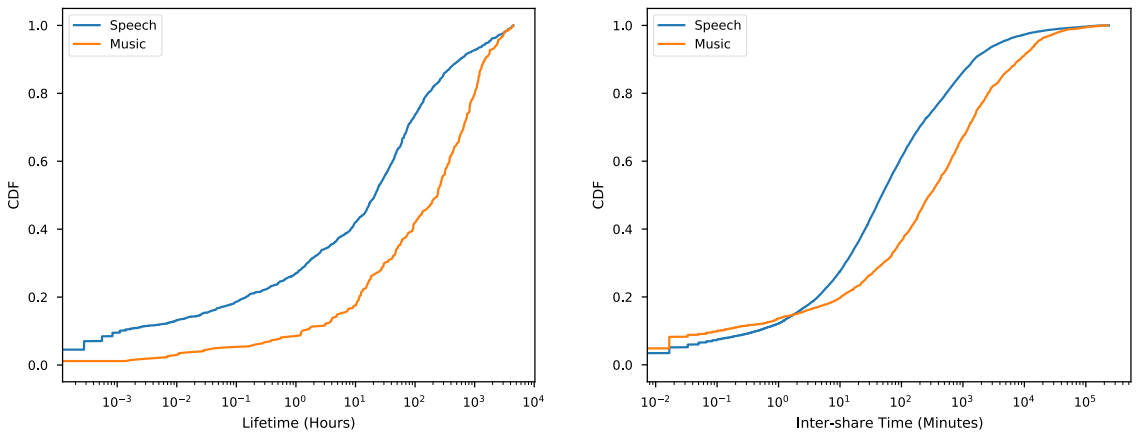
2. They have many psychological attributes that catch people's attention and have a direct impact on our emotions, such as the use of negative words, or attributes regarding our future, as seen on the psychological attribute analysis using LIWC in Section 4.1.2 and even in the response from the interviews conducted in Section 4.2;

3. Audios with misinformation often are accompanied by many different characteristics, which make them more engaging, such as "sources" that try to back their story. They also try to engage the listener in actions (e.g., re-sharing) as seen in Section 4.2.

These observed characteristics of audios with misinformation may contribute to their great attractiveness and virality. An interesting avenue of future work is to explore the greater presence of these properties in the design of methods to detect misinformation and mitigate its harmful impact.

4.3.3 Speech versus Music and Gender Differences

Finally, we analyze the differences in propagation dynamics between audios with music and speech. Figure 4.8 shows the distributions of lifetimes and inter-share times for audios with speech and music as their primary type of content. Audios with music have a significantly higher lifetime than speech audios. While 50% of speech content has a lifetime of around one day, half of the music audio has a lifetime of at least ten days. The inter-share time of music audios is also higher. Around 50% of speech audios have an inter-share time of 1 hour, while half of the audios containing music have an inter-share time of 5 hours. One explanation for the inter-share time is the higher effort to listen to them, as they often are lengthier than speech audios, as the average length of speech, audio is 122 seconds, while music is on average 182 seconds (statistically different according to a t-test with p-value ≤ 0.05).



(a) Lifetime (b) Inter-share Time

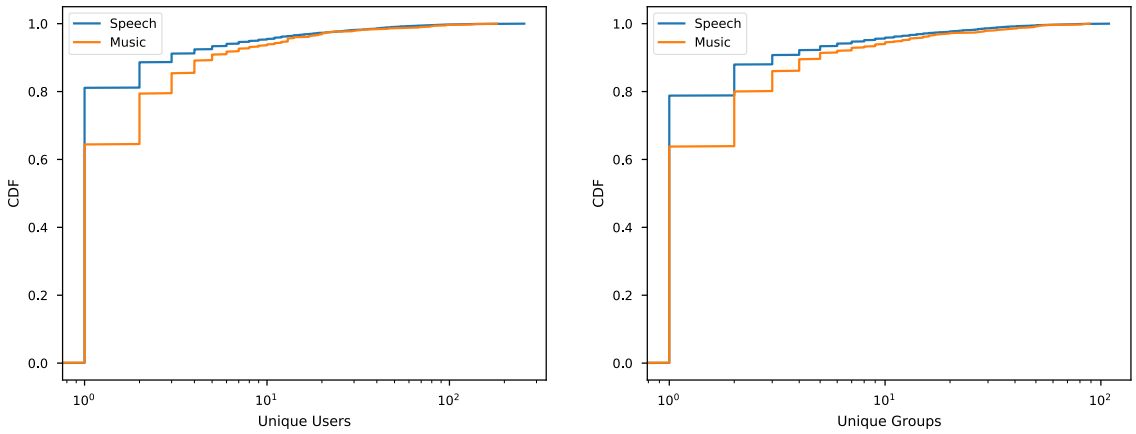
Figure 4.8. Distributions of lifetimes and inter-share times of audio messages in audios with speech and music

Table 4.7 shows average and standard deviations of the numbers of groups, sharing users and times shared per audio message of each kind. Figure 4.9 shows corresponding distributions. In all three distributions, we notice that the musical audios are the most spread, appearing in more groups, being shared by more users, and having an average

share count greater than the speech audios. In terms of averages (Table 4.7), audios with music were shared by 28% more users, in 43% more groups and 30% more times. We note that most music audios contained some political content, which may justify the greater attractiveness. The top ten most shared musical audio, which were shared more than 80 times each (the most shared audio was shared 223 times), were political propaganda for the presidential candidate Jair Bolsonaro.

Table 4.7. Speech vs. music spreading

Type	Analysis	Mean ± Standard Deviation
Speech	Number of groups per message	3.39 ± 0.08
	Number of sharing users per message	3.01 ± 0.13
	Number of times shared per message	2.53 ± 0.15
Music	Number of groups per message	4.88 ± 0.28
	Number of sharing users per message	3.88 ± 0.45
	Number of times shared per message	3.29 ± 0.57



(a) Unique Users per Message

(b) Unique Groups per Message

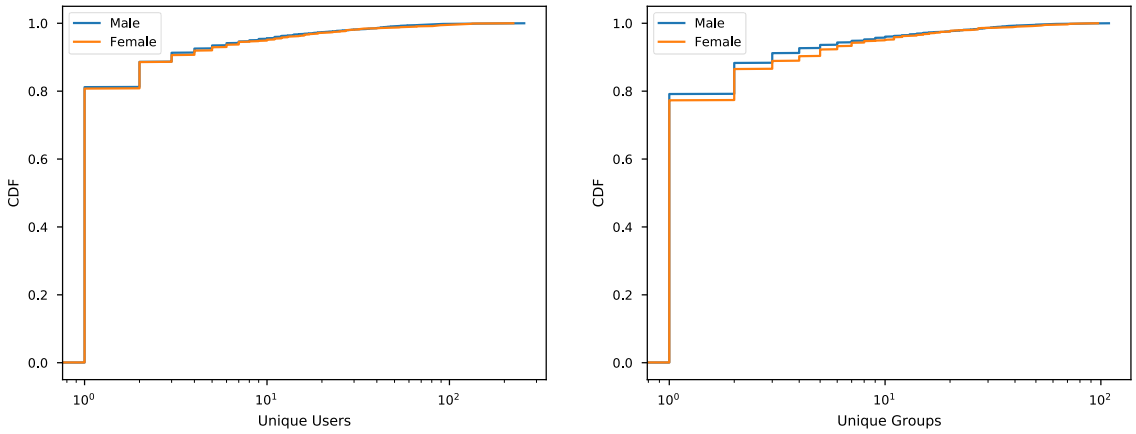
Figure 4.9. Distribution of Unique Users and Unique Groups per Message for Speech versus Music

We also analyzed the propagation dynamics of audios with male and female speakers. Table 4.8 shows the averages and standard deviation of the number of sharing users, the number of groups, and the number of times shared. Figure 4.10 shows the distributions of the number of sharing users and the number of groups. Despite some differences in average values, suggesting that audios with primarily female speakers tend to have a greater reach, we found no statistical difference between the three pairs of distributions (according to a Kolmogorov-Smirnov test). Similarly, according to a

t-test, the averages cannot be considered statistically different. There were also no significant differences in the lifetime and inter-share time distributions. Thus, gender does not seem to play a significant role in the propagation dynamics of audio content in WhatsApp.

Table 4.8. Male vs. Female spreading

Type	Analysis	Mean ± Standard Deviation
Male	Number of groups per message	2.48 ± 0.09
	Number of sharing users per message	2.93 ± 0.14
	Number of times shared per message	3.29 ± 0.16
Female	Number of groups per message	2.76 ± 0.21
	Number of sharing users per message	3.33 ± 0.36
	Number of times shared per message	3.81 ± 0.41



(a) Unique Users per Message

(b) Unique Groups per Message

Figure 4.10. Distribution of Unique Users and Unique Groups per Message for Male versus Female speakers

4.4 Summary

In this chapter, we looked over the content and propagation dynamics of the audios messages from the collected WhatsApp dataset. We started by looking at the audio messages' content by using the transcriptions from the speech recognition phase to perform a topic analysis and collected psychological linguistic features based on an LIWC dictionary. We were able to identify eight topics of discussion in the audios, where four topics were directly related to politics and had the most misinformation

related to them. One topic was highly related to religious words, and the other three were more related to chatter and contained more general words. From the LIWC attributes, we identified that audios with misinformation had a higher presence of the attributes *negemo* (negative emotion) and the *insight* attribute (characterized by words such as "attention", "warning" and "listen"), carried more often phrases in the future tense and often talked directly to the listener by using words such as "you". Comparing to other studies where the authors analyzed misinformation in textual content, the *insight* attribute was present in both types of content, but textual content was more aimed at aggregating the community, using words such as "we" and textual content often used the present tense, indicating different types of approach depending on the media being used (text or audio).

We then proceeded to a qualitative analysis, composed of an interview and an online survey with volunteers. The purpose of this step was to deepen our knowledge about the audio messages, gathering the perception of selected volunteers on the audio's content and potential feelings toward the speaker's voice. We noted that audios with misinformation tend to make the listener feel negative emotions, such as sadness, anger, and disgust, linking to our previous LIWC analysis, where similar attributes were found in audio messages with misinformation. Volunteers often noted that audios with misinformation tried to back their claims by citing some sources. However, the volunteers often saw these sources as unreliable and did not make the information more believable. The speaker's tone from the audios with misinformation was considered less *friendly* and less *natural* than audios with unchecked content. We also noted that audios with misinformation are often accompanied by instructions for the listener, such as sharing the audio to their group of friends or relatives.

We finish the discussion by looking at the propagation dynamics, such as lifetime and inter-share times, and also metrics such as the number of groups per message, number of sharing users per message, and number of times shared per message. Firstly, we compare the lifetime and inter-share time between two major events in Brazil, the Trucker strike and Electoral Campaign. The propagation behavior between these periods is relatively similar to each other. We also compared these metrics with textual and image content propagation from the studies by Resende et al. [2019a,b]. Audio messages tend to spread more slowly but remain for a shorter period in the system. We then shift our focus to the propagation dynamics of audio messages with misinformation versus unchecked content. Audio messages with misinformation tend to last three weeks more and spread somewhat quicker than unchecked content. These numbers represent a significant increase when looking at the propagation of image content with misinformation but had similar behavior to textual messages with misinformation.

Lastly, we look at the propagation of speech versus music. Audios with music stayed significantly longer in the system but also had a higher inter-share time, spreading more slowly. Many of the audios containing music were related to electoral campaign jingles. The gender of the speaker did not seem to play a significant role in the propagation dynamics of audio content in WhatsApp.

Chapter 5

Conclusions and Future Work

In this master thesis, we looked into audio communication in publicly accessible WhatsApp groups. WhatsApp is one of the primary forms of communication in many countries, such as Brazil and its usage has raised several concerns in the past few years regarding misinformation spread. The application has some key features that make it stand out from other platforms: end-to-end encryption, making the messages accessible by only those involved in the conversation; creation of groups (which can be made publicly accessible by sharing an invite link at large), and finally; there are features for quickly sharing messages, such as forwarding messages to other groups and users, or by creating broadcast lists where a single message can be sent to multiple groups at once. These features, combined with the high market penetration in some countries, brought to light some worrisome behaviors, specifically those related to the dissemination of misinformation.

Given that the spread of misinformation in WhatsApp groups is rapidly increasing and is negatively affecting many recent discussions, it becomes of interest to understand how these messages propagate in the app and what the unique characteristics of these messages are to get a sense of how the app is currently being exploited to boost this kind of content. Recent studies looked at the propagation of textual and image content in WhatsApp, but to our knowledge, no study focused on audio content.

In that context, our primary aim was to get a better understanding of how audio messages are used in publicly accessible WhatsApp groups. We first focused on understanding the characteristics of these audio messages in terms of content properties and propagation dynamics while also looking at the differences to prior findings for other types of content (textual and image). We also looked at the introspect properties of audio content, such as the gender of the speaker, checking how these properties correlate with propagation dynamics. Finally, we also analyzed how these properties differ

between audio messages carrying previously checked misinformation and unchecked content.

We started by proposing a pipeline to analyze the audio messages shared in publicly accessible WhatsApp groups composed of seven steps: (1) pre-processing; (2) similarity detection, to group audios with equivalent content; (3) speech recognition to transcribe the audios; (4) misinformation detection based on the audio transcription and fact-checked articles from fact-checking agencies; (5) audio type categorization (speech versus music and gender classification); (6) a qualitative analysis and finally; (7) content and propagation analysis. In our analysis we also contrasted the differences between audios with previously checked misinformation and unchecked content.

To understand the content of the audio messages, we relied on two strategies: a topic analysis using LDA and the extraction of psychological linguistic features using LIWC. Regarding the topics of discussion, we identified eight main topics discussed in the audio messages. Four topics were directly related to politics (were linked to political words, e.g., "Military", "Haddad", "Bolsonaro", "PT") and had the largest fraction of misinformation related to them. One topic was highly related to religious words, and the other three were more related to chatter. From the LIWC attributes, we identified that audios with misinformation had a higher presence of negative emotions, and used more words related to the *insight* attribute, such as "attention", "warning" and "listen". They also used more phrases in the future tense, and talked directly to the listener by using words such as "you". Prior analyses of textual content shared on WhatsApp found the frequent presence of terms that aggregate the community, such as "we", and verbs often in the present tense. Thus, our present findings, different from those prior ones, indicates different types of approach depending on the media being used (text or audio). Negative emotion terms are not quite as present in textual content as found in audio content.

We conducted a qualitative analysis based on two phases: an interview and an online survey. The primary objective was to deepen our knowledge about the audio messages, gathering the perception of selected volunteers on the audio's content and potential feelings the speaker's voice triggered, analyzing audios with misinformation and with unchecked content separately. One key result from the qualitative analysis is that audios with misinformation tend to more often make the listener feel negative emotions, such as sadness, anger, and disgust. The volunteers also noted that the audios with misinformation often tried to back their claims by citing some sources. However, the volunteers often saw these sources as unreliable and did not believe they made the information more believable. The speaker's tone from the audios with misinformation was considered less *friendly* and less *natural* than audios with unchecked

content. Lastly, volunteers also noted that audios with misinformation carried some instruction for the listener, such as sharing the audio with other groups. These "call to actions" were not seen as often in audios with unchecked content.

Finally, we looked into how these audios propagated in these groups by looking at lifetime and inter-share time metrics. Overall, 60% of audios are re-shared within 3 hours, and 20% are re-shared within 6 minutes. Comparing these results with prior analyses, we noticed that audios tend to spread more slowly and remain for shorter periods in the system than textual content. This could reflect a consequence of the greater effort the user has to put to listen to an audio message compared to reading a text. When comparing audios with misinformation and unchecked content, we were able to see that audios with misinformation spread quicker than unchecked audios and last significantly longer in the network: 75% of audios with unchecked content tend to last at most seven days, while 75% of audios with misinformation last up to 31 days.

We observed that misinformation audios appear in more groups, are shared by more users, and have overall more shares than unchecked content. This could be explained by many factors, such as being targeted for incredibly relevant topics to the current political scenario that they appear in, having many psychological attributes that catch people's attention, and directly impacting the listeners' emotions and being more engaging. We also noted that audios containing music had a higher reach than speech audios, appearing in more groups and shared by more users. They also had a considerably longer lifetime (while half of the speech audios had a lifetime of a single day, half of the music audios lasted for ten days). They also had a longer inter-share time, which could be explained by the higher effort to listen to them, as they are longer on average. Usually, these audios contained electoral campaign jingles. The gender of the speaker did not seem to play a significant role in the propagation dynamics of audio content in WhatsApp.

Overall, we confirmed that audio communication is widely used in Brazil and was extensively used in the Electoral Campaign and the Trucker Driver Strike, two major events that took place in Brazil in 2018. WhatsApp is a massive communication tool used by millions of people who can and are influencing people worldwide. Due to its very own nature of being end-to-end encrypted, these communications are secured within each person's phone, and the information running in the network is mostly unknown. Understanding how these tools are being used and if they are being targeted for massive misinformation propagation is an essential first step in planning ways to create awareness or methods for stopping this undesired effect without necessarily violating the user's privacy. In this work, we presented how audio-based communication is used in WhatsApp and that it is, in fact, a tool for propagating misinformation

across the network, just as text and images are. To our knowledge, this study is one of the first to tackle audio communication in WhatsApp. It revealed that this form of communication follows different patterns from text and images content, especially when it comes to misinformation; therefore, this work's main findings complement the literature.

A possible direction for the future consists of expanding our analysis to account for audios shared across many years, looking into how these properties behave across time, and possibly detecting seasonal events. Another direction to be pursued is to expand our misinformation detection pipeline to reliably and automatically detect audios with misinformation, thus expanding our current misinformation analysis to a larger quantity of audio files. Moreover, we would like to expand the analysis to more than one country, such as countries with different levels of education, looking at similarities and differences that can emerge. More broadly, expanding the analysis of audio communication to other social platforms and assessing how audio communication is used in each one of them compared to WhatsApp.

Bibliography

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

Antoniadis, S., Litou, I., and Kalogeraki, V. (2015). A model for identifying misinformation in online social networks. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 473--482. Springer.

Baeth, M. J. and Aktas, M. S. (2019). Detecting misinformation in social networks using provenance data. *Concurrency and Computation: Practice and Experience*, 31(3):e4793.

Bartsch, M. and Wakefield, G. (2005). Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104. ISSN 1520-9210.

Bartsch, M. A. and Wakefield, G. H. (2005). Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on multimedia*, 7(1):96--104.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359--370. Seattle, WA.

Bessi, A. and Ferrara, E. (2016). Social bots distort the 2016 us presidential election online discussion. *First Monday*, 21(11-7).

Bhatia, R., Srivastava, S., Bhatia, V., and Singh, M. (2018). Analysis of audio features for music representation. In *2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 261–266. ISSN .

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993--1022.

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., and Dredze, M. (2018). Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American journal of public health*, 108(10):1378--1384.

Bursztyn, V. S. and Birnbaum, L. (2019). Thousands of small, constant rallies: A large-scale analysis of partisan whatsapp groups. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, page 484–488, New York, NY, USA. Association for Computing Machinery.

Caetano, J. A., Magno, G., Gonçalves, M. A., Almeida, J. M., Marques-Neto, H. T., and Almeida, V. A. F. (2019). Characterizing attention cascades in whatsapp groups. *CoRR*, abs/1905.00825.

Cano, P., Batle, E., Kalker, T., and Haitsma, J. (2002). A review of algorithms for audio fingerprinting. In *2002 IEEE Workshop on Multimedia Signal Processing.*, pages 169--173. IEEE.

Conti, M., Lain, D., Lazzeretti, R., Lovisotto, G., and Quattrociocchi, W. (2017). It's always april fools' day!: On the difficulty of social network misinformation classification via propagation features. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6. ISSN .

Cunningham, S., Ridley, H., Weinel, J., and Picking, R. (2020). Supervised machine learning for audio emotion recognition. *Personal and Ubiquitous Computing*.

de Assis, C. (2019). Cresce uso de instagram e whatsapp para consumo de notícias online em argentina, brasil, chile e méxico, aponta relatório. https://knightcenter.utexas.edu/pt-br/blog/00-20990-cresce-uso-de-instagram-e-whatsapp-para-consumo-de-noticias-online-em-argentina-brasil. Accessed in 2019-10-06.

Denny, M. and Spirling, A. (2017). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *When It Misleads, and What to Do about It (September 27, 2017)*.

DiCicco-Bloom, B. and Crabtree, B. F. (2006). The qualitative research interview. *Medical education*, 40(4):314--321.

Doukhan, D., Carrive, J., Vallet, F., Larcher, A., and Meignier, S. (2018). An open-source speaker gender detection framework for monitoring gender equality. In *IEEE ICASSP*.

Ellis, D. P., Whitman, B., and Porter, A. (2011). Echoprint: An open music identification service. *International Society for Music Information Retrieval*.

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Commun. ACM*, 59(7):96--104. ISSN 0001-0782.

Ferreira, C. D. (2019). Facebook chega a 127 milhões de usuários mensais no brasil. https://www1.folha.uol.com.br/tec/2018/07/facebook-chega-a-127-milhoes-de-usuarios-mensais-no-brasil.shtml. Accessed in 2019-09-01.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Fourney, A., Racz, M. Z., Ranade, G., Mobius, M., and Horvitz, E. (2017). Geographic and temporal trends in fake news consumption during the 2016 us presidential election. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2071--2074. ACM.

Fraser, M. T. D. and Gondim, S. M. G. (2004). Da fala do outro ao texto negociado: discussões sobre a entrevista na pesquisa qualitativa. *Paidéia (Ribeirão Preto)*, 14(28):139--152.

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., and Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214--224.

Giraudel, A., Carré, M., Mapelli, V., Kahn, J., Galibert, O., and Quintard, L. (2012). The repere corpus: a multimodal corpus for person recognition. In *LREC*, pages 1102--1107.

Goldberg, Y. and Levy, O. (2014). word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Graves, L. (2013). *Deciding what's true: Fact-checking journalism and the new ecology of news*. PhD dissertation, Columbia University.

Han, K. J., Prieto, R., Wu, K., and Ma, T. (2019). State-of-the-art speech recognition using multi-stream self-attention with dilated 1d convolutions. *arXiv preprint arXiv:1910.00716*.

Herchonvicz, A. L., Franco, C. R., and Jasinski, M. G. (2019). A comparison of cloud-based speech recognition engines. *Anais do Computer on the Beach*, pages 366--375.

Iqbal, M. (2019). Whatsapp revenue and usage statistics. https://www.businessofapps.com/data/whatsapp-statistics/. Accessed in 2019-09-01.

Jang, D., Yoo, C. D., Lee, S., Kim, S., and Kalker, T. (2009). Pairwise boosted audio fingerprint. *IEEE Transactions on Information Forensics and Security*, 4(4):995–1004. ISSN 1556-6013.

Jang, D., Yoo, C. D., Lee, S., Kim, S., and Kalker, T. (2009). Pairwise boosted audio fingerprint. *IEEE transactions on information forensics and security*, 4(4):995--1004.

Jensen, J. H., Christensen, M. G., Murthi, M. N., and Jensen, S. H. (2006). Evaluation of mfcc estimation techniques for music similarity. In *2006 14th European Signal Processing Conference*, pages 1--5. IEEE.

Juang, B. H. and Rabiner, L. R. (1991). Hidden markov models for speech recognition. *Technometrics*, 33(3):251--272.

Kotti, M. and Kotropoulos, C. (2008). Gender classification in two emotional speech databases. In *2008 19th International Conference on Pattern Recognition*, pages 1--4. IEEE.

Kumar, K. K. and Geethakumari, G. (2014). Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences*, 4(1):14.

Larson, M., de Jong, F., Kraaij, W., and Renals, S. (2012a). Special issue on searching speech. *ACM Transactions on Information Systems (TOIS)*, 30(3):1--2.

Larson, M., Jones, G. J., et al. (2012b). Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends® in Information Retrieval*, 5(4–5):235--422.

Logan, B. and Salomon, A. (2001). A music similarity function based on signal analysis. *International Conference on Multimedia and Expo*.

Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., and Gay, G. (2006). The influence of task and gender on search and evaluation behavior using google. *Information processing & management*, 42(4):1123--1131.

Loubak, A. L. and Achilles, R. (2019). Whatsapp admite envio ilegal de mensagens em massa nas eleições 2018. https://www.techtudo.com.br/noticias/2019/10/whatsapp-admite-envio-ilegal-de-mensagens-nas-eleicoes-2018.ghtml. Acessed in 2019-10-12.

Lüscher, C., Beck, E., Irie, K., Kitza, M., Michel, W., Zeyer, A., Schlüter, R., and Ney, H. (2019). Rwth asr systems for librispeech: Hybrid vs attention-w/o data augmentation. *arXiv preprint arXiv:1905.03072*.

Martins, A. (2018). Na web, 12 milhões difundem fake news políticas. https://politica.estadao.com.br/noticias/geral,na-web-12-milhoes-difundem-fake-news-politicas,70002004235. Accessed in 2019-09-1.

McKinney, M. and Breebaart, J. (2003). Features for audio and music classification.

Meinedo, H. and Trancoso, I. (2010). Age and gender classification using fusion of acoustic and prosodic features. In *Eleventh Annual Conference of the International Speech Communication Association*.

Meixler, E. (2018). Five killed in latest mob attack after rumors on social media. here's what to know about india's whatsapp murders. https://time.com/5329030/india-whatsapp-murders-mob-false-rumors/. Accessed in 2019-09-1.

Melo, P., Messias, J., Resende, G., Garimella, K., Almeida, J., and Benevenuto, F. (2019a). Whatsapp monitor: A fact-checking system for whatsapp. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 676--677.

Melo, P., Vieira, C. C., Garimella, K., de Melo, P. O. S. V., and Benevenuto, F. (2019b). Can whatsapp counter misinformation by limiting message forwarding?

Menczer, F. (2016). The spread of misinformation in social media. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 717--717.

Mueller, W., Silva, T. H., Almeida, J. M., and Loureiro, A. A. (2017). Gender matters! analyzing global cultural gender preferences for venues using social sensing. *EPJ Data Science*, 6(1):5.

Müller, M. (2007). Dynamic time warping. *Information retrieval for music and motion*, pages 69--84.

Ooi, C. S., Seng, K. P., Ang, L.-M., and Chew, L. W. (2014). A new approach of audio emotion recognition. *Expert systems with applications*, 41(13):5858--5869.

Ortega, J. D., Senoussaoui, M., Granger, E., Pedersoli, M., Cardinal, P., and Koerich, A. L. (2019). Multimodal fusion with deep neural networks for audio-video emotion recognition. *arXiv preprint arXiv:1907.03196*.

Pennebaker, J., Boyd, R., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532--1543.

Picone, J. W. (1993). Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1215--1247.

Porter, A. (2013). *Evaluating musical fingerprinting systems*. PhD dissertation, McGill University Libraries.

Qazvinian, V., Rosengren, E., Radev, D. R., and Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1589--1599, Stroudsburg, PA, USA. Association for Computational Linguistics.

Quintanilha, I. M. (2017). *End-to-end speech recognition applied to brazilian portuguese using deep learning*. PhD dissertation, MSc dissertation, PEE/COPPE, Federal University of Rio de Janeiro, Rio de ....

Reis, J. C. S., de Freitas Melo, P., Garimella, K., and Benevenuto, F. (2020). Can whatsapp benefit from debunked fact-checked stories to reduce misinformation?

Resende, G., Melo, P., Reis, J., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019a). Analyzing textual (mis)information shared in whatsapp groups. In *WebSci '19*.

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019b). (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *The World Wide Web Conference*, WWW '19, page 818–828, New York, NY, USA. Association for Computing Machinery.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399--408.

Roozenbeek, J. and Van Der Linden, S. (2019). The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research*, 22(5):570--580.

Salmon, F. and Vallet, F. (2014). An effortless way to create large-scale datasets for famous speakers. In *LREC*, pages 348--352.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513--523.

Salvador, S. and Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561--580.

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L., et al. (2017). English conversational telephone speech recognition by humans and machines. *arXiv preprint arXiv:1703.02136*.

Seufert, M., Hoßfeld, T., Schwind, A., Burger, V., and Tran-Gia, P. (2016). Group-based communication in whatsapp. In *2016 IFIP networking conference (IFIP networking) and workshops*, pages 536--541. IEEE.

Sherman, L. E., Michikyan, M., and Greenfield, P. M. (2013). The effects of text, audio, video, and in-person communication on bonding between friends. *Cyberpsychology: Journal of psychosocial research on cyberspace*, 7(2).

Tang, C., Ross, K., Saxena, N., and Chen, R. (2011). What's in a name: A study of names, gender inference, and gender behavior in facebook. In *International Conference on Database Systems for Advanced Applications*, pages 344--356. Springer.

Tardáguila, C., Benevenuto, F., and Ortellado, P. (2018). Fake news is poisoning brazilian politics. whatsapp can stop it. https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html. Accessed in 2019-10-06.

Vallet, F., Uro, J., Andriamakaoly, J., Nabi, H., Derval, M., and Carrive, J. (2016). Speech trax: A bottom to the top approach for speaker tracking and indexing in an archiving context. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2011--2016.

Wagner, M. C. and Boczkowski, P. J. (2019). The reception of fake news: The interpretations and practices that shape the consumption of perceived misinformation. *Digital Journalism*, 7(7):870--885.

Wang, A. et al. (2003). An industrial strength audio search algorithm. In *Ismir*, volume 2003, pages 7--13. Washington, DC.

Yang, L., Wang, Y., Dunne, D., Sobolev, M., Naaman, M., and Estrin, D. (2019). More than just words: Modeling non-textual characteristics of podcasts. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 276--284.

Yoon, S., Byun, S., and Jung, K. (2018). Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112--118. IEEE.

Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43--52.

Zhou, X. and Zafarani, R. (2018). Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*.