MASTER IN COGNITIVE SCIENCE AND LANGUAGE

MASTER THESIS

September 2024

Exploring the role of prosodic information as a modality in hate speech detection

by Alejandra Duque Maldonado

Under the supervision of:

Mireia Farrús Cabeceran

Abstract

The multimodal detection of hate speech has been a trending topic across different disciplines in the recent years. New approaches make use of multimodal techniques to target and mitigate toxic behaviour. Current proposals, despite relying on various modalities, tend to prioritize the use of text in monolingual settings, typically with English. In this thesis, we want to avoid using textual data, and rather focus on the audio modality to see if its properties can help us target toxic speech. Given that within speech prosody we could target possible cues related to indicators of hate speech (e.g. emotional speech), our aim is to test the effectiveness of predicting toxic speech based on prosodic features obtained from the audio.

We used two different classification methods to test whether the use of these audio properties was satisfactory. These algorithms were trained on a database of YouTube videos in Spanish, within which there were examples of hate speech towards gender and equality speeches. These examples were processed to work only with the speech signal, which naturally reflected prosodic properties.

In the scope of our application, we saw that our traditional machine learning approach suggests it is possible to detect hate speech based on prosodic information. By using frame-wise information of the audio, we have seen that it is possible to classify hate speech automatically. Our proposal opens the door for future works to continue testing the effectiveness of including this information, and encourages future proposals to merge prosodic information with other modalities.

Acknowledgements

First of all, I want to thank my supervisor Mireia Farrús for guiding me through the process of writing this thesis. Her advice has been very valuable for me, and I have learned a lot from her.

I would like to thank the members of the Centre de Llenguatge i Computació (CLiC) research group, especially Laura de Grazia and Mariona Taulé, for letting me contribute to their research and advising me during this process. Moreover, I would like to thank the Servei de Tecnologia Lingüística (STeL) group for their support, and for helping me use tools to carry out this work more efficiently.

On a personal note, I would like to thank my family, both here and in Venezuela. They have been my greatest support in all my academic journey and, despite the distance, they have always been there no matter the circumstances. I want to thank my partner, for supporting and advising me everyday when everything seemed more difficult. He has been a great support to me, and has helped me find clarity when I needed it most. Lastly, I want to thank my colleagues. For their friendship, for making this master's time so enjoyable and for their unconditional support in every class.

Contents

1 Introduction **6**

1.1 Motivation and Problem statement 6

1.2 Structure of the thesis 7

2 Literature review **8**

2.1 Theoretical framework 8

2.2 Related works 10

3 Methodology **12**

3.1 Materials 12

3.1.1 Collection 12

3.1.2 Annotation 13

3.2 Preprocessing of data 14

3.2.1 Data normalization and adaptation 14

3.2.2 Feature extraction 15

3.3 Exploratory analysis 18

3.3.1 T-test 18

3.3.2 Support Vectors Machine (SVM) 19

3.4 Classifiers 21

3.4.1 Support Vector Classifier (SVC) 21

3.4.2 Convolutional neural networks (CNN) 22

4 Results **24**

4.1 Exploratory analysis 24

4.2 Classifiers 26

4.3 Discussion 28

5 Conclusions and Future Work **29**

6 Appendices **29**

A Key search terms **29**

B Playlists used in data collection **33**

C Annotation criteria for labelling hate speech **33**

References **34**

1 Introduction

Nowadays, the extended use of social media has caused a great impact in society. People are constantly in contact with networks where they can instantly access information, interact with other users, and broadcast their views and opinions about any topic. However, the increased use of social media has opened the doors to the spread of hateful content and cyberbullying in these platforms. This is the reason why current technologies have attempted to find innovative solutions to effectively target toxic behaviours.

At the intersection between Natural Language Processing (NLP) and automated detection, targeting hate speech (HS) has been recognized as a challenging task that calls for new approaches and further research. Given the multimodal nature of today's content, automated detection must account for the numerous features found in media. It should also account for the linguistic complexity behind formulating and decoding toxic messages.

Considering that human speech conveys meaning beyond textual content, we want to determine whether using intonational information, such as prosody, could be relevant in determining HS. Taking into account the advances made in this field, we explore a relatively new approach that comprises the use of the properties of speech in the detection of HS. Therefore, our final goal is to determine whether the prosodic information could potentially be a predictor of hateful content.

1.1 Motivation and Problem statement

Toxic discourse is complex and requires many considerations when determining what signals its presence. Linguistically, formulating and decoding these messages depends heavily on the semantic and pragmatic knowledge of the interlocutors. In other words, interlocutors must be aware of the semantic and pragmatic context to formulate and understand hate messages or attacks. However, it should also be noted that human language not only makes use of morphosyntactic information to convey messages, there are complementary devices, such as intonation and gestures, used to convey secondary messages that enrich what is said. In HS, these explicit and implicit devices are used in order to send complex messages. These devices can be used strategically either to make direct attacks, to suggest implicitly coded hate messages or even to emphasise offensive behaviours and messages.

In automated detection, representing the complexity of hate messages is rather challenging. Natural information, such as speech, can hardly be represented with a single modality, which sparks the necessity of having multiple sources of data that altogether can introduce enough information to represent the intricacy of the source (Lahat, Adali, & Jutten, 2015). Data fusion methods have found a way to process the multimodal data present in social media, and suggests that media complexity can be represented closely enough to targeting some forms of HS (e.g. in image-text media). However, the majority of current approaches to multimodal HS detection are focused on text, making it the main source of reliable information, and using additional modalities for complementing what is extracted from the text.

In this task, we want to move away from this textual focus and give priority to the properties found in other modalities, which could potentially be as informative of HS. Considering that humans use intonation for complementing most of the messages we communicate, and given that intonation changes are quantifiable and visible in the speech signal, we want to test the hyphotesis that intonational information could be used to potentially represent this type of discourse. We hypothesise that if HS has consistent characteristics, the prosody of speech could become informative of its presence. If this is the case, we expect this information can be extracted and included in multimodal tasks for complementing other modalities, and serve as a disambiguator in those cases where other modalities might not be enough for determining HS.

On this note, this dissertation will focus on the exploration of how informative prosodic information is when determining HS, and whether it can potentially be used in automated detection for predicting hateful content.

1.2 Structure of the thesis

This thesis is structured as follows. Section 2 is dedicated to understanding the definition of prosody and its contribution to human communication. Similarly, we will review previous works on multimodal detection of HS, specifically those that have similarities to our approach and the state-of-the-art. Section 3 describes the methodological process that lead to our final results. This will delineate the origins of our data (its collection and annotation), preprocessing, feature extraction and implementation of the different classification models selected. Section 4 displays

the results and discussion of stemming from the application of our experiments. Lastly, Section 5 delineates the final conclusions of our analysis, and states open questions that can be undertaken in future works.

2 Literature review

2.1 Theoretical framework

In human communication, information is not solely conveyed through words and linguistic information, there are extralinguistic resources that contribute to the meaning of *what* is being said and *why* we communicate it. Prosody refers to the phonetic and phonological properties of speech that bring musicality to linguistic units. Its realization involves both segmental and suprasegmental features of speech, which conveys linguistic information —explicitly represented by discrete symbols—and paralinguistic information, non-verbal cues that may carry information about the message and that are normally interpreted by inference (i.e. "body language" as gestures, eye-movements or sounds). How these items are organized within an utterance (or linguistic unit) depends on the syntactic function and semantic information of what we want to communicate. In other words, how the speaker rhythmically places these units conveys information about the semantic content and purpose of the utterance (Fujisaki (1997); Gussenhoven (2002); Ephratt (2011)).

Prosodic markers are the resources we naturally use to convey semantic meaning of morphosyntactic constituents or inform about the illocutionary act of an utterance. We use them to communicate what is not explicitly stated, and what listeners must infer based on linguistic and paralinguistic evidence (Bryant, 2010). These markers can manifest in intonation through pitch variations, intensity changes and phrasing (i.e. grouping of utterances or pauses) (Gussenhoven, 2002). They can also be observed empirically, as these are parameters that reflect in the speech signal.

Prosodic markers can be used to disambiguate meaning between words (e.g. by marking prosodic accent), inform about the emotional state or behaviour of the speaker (e.g. being friendly or showing anger), signal turn-taking in a conversation, mark emphasis, or inform about the communicative function of an utterance (e.g. raising intonation at the end to signal a question),

among other uses. Depending on the pragmatic context, these uses vary, but overall they are used in order to signal implicit information within an utterance. In a Gricean sense, intentions drive speakers' behaviors (e.g. utterances) whose sole function is to have an effect on the addressee in virtue of having their intention recognized (Hellbernd & Sammler, 2016). This implies that meaning is not necessarily fully reflected in the lexical content of the utterance, there can be extralinguistic content, that manifests through prosodic signals, which the recipient can use to infer the speaker's intentions.

In the context of HS, some indicators of its presence are seen in ambiguous content of statements, irony, sarcasm, rhetorical devices, mockery and emotional speech (Papcunová et al., 2021). Although these items are not exclusive to HS, they are resources often used to communicate explicit or implicit attacks, and which realizations often have some characteristic musicality. Considering that we use prosody to communicate affective states (De Moraes, 2011), we could expect that some indicators manifest in the speech signal and can be used as potential cues to target toxic speech[1].

It is worth noting that not all forms of HS could manifest through prosodic cues. There are forms of HS that seamlessly integrate in discourse, such as dogwhistles—terms that send one message to an outgroup whilst sending a second message to an ingroup, whose meaning can be deduced by inference and typically does not carry extralinguistic information, e.g. emotional cues (Henderson & McCready, 2018). The existence of such phenomena makes us realise that some linguistic indicators to HS are not always present, in which case, messages could only be decoded from pragmatic knowledge. Regardless of this fact, although there are cases in which HS is meant to be disguised, we expect that prosodic patterns might be able to signal the presence of other indicators commonly seen in HS —such as the use of irony—and, therefore, allow to identify hate messages through other non-linguistic indicators.

On this basis, the aim of our approach is to use prosodic information in favour of highlighting these suprasegmental features, and use them to obtain information about *how* a message is being communicated. In a multimodal setting, we expect this information about the speech can disambiguate or directly signal whether a message can contain hatred, in case the information from other modalities (e.g. visual or textual information) is not conclusive.

[1]Depending on the definition of "hate speech'' that is assumed, the indicators of its presence might vary. In Section 3.1.2 we provide details on the definition and indicators we adopt to identify HS.

2.2 Related works

Systematic reviews as the one by Chhabra and Vishwakarma (2023) bring to light the current state of multimodal detection in this field. Traditionally, the main focus has been placed on the textual modality in monolingual settings, normally in the English language. However, the trend has recently shifted towards multimodal and multilingual approaches using machine learning or neural networks. Although now the focus is placed on multimodality, the majority of automated detection still relies on text, including audio-based detection; where the audio is used mainly used for extracting textual information from the soundwave, rather than using the insights of the properties of the audio.

Nevertheless, earlier research works show that extracting prosodic-related features from the soundwave can help in the automated classification of different types of speech. For instance, authors like Sato, Mitsukura, Fukumi, and Akamatsu (2001) and Luengo, Navas, Hernáez, and Sánchez (2005) used pitch, energy and time-series data for classifying different emotions by using neural networks and SVM. Although this suggests prosody can be used to classify types of emotional speech, in HS we are interested in looking beyond emotions, as there are other types of indicators —such as figures of speech, irony or questioning—that can manifest in the speech signal with different characteristics. As far as of speech acts are concerned, past works have also used acoustic information to try to classify them. Bryant (2010) and Hellbernd and Sammler (2016) have shown through the analysis of spontaneous speech that different speech acts (e.g. irony, criticism, doubt, naming, suggestion, warning...) can be separated by using acoustic features, which are independent from lexical content. These works show that possible indicators of HS can be identified by their acoustic properties. In this dissertation, we want to test whether these properties can highlight HS in automatic detection by looking at the acoustic features of the audio.

In HS studies, we are beginning to see literature related to prosody in HS perception and production. Research by author Niebuhr, O. has explored on the role of prosody in the production and perception of HS in Germanic languages. In preliminary studies, Neitsch and Niebuhr (2019) have observed the acoustic manifestation of spoken HS. Here, they found there are context-independent phonetic patterns that characterize HS; for instance, an increase in phonetic effort (i.e higher HNR

and Hammarberg-index values[2]) as a consequence of wanting to communicate messages clearly. Later on, Neitsch and Niebuhr (2020) continued to examine german HS tokens and how they are perceived, finding that prosody has an impact in how strongly HS is perceived (e.g. low F_0, breathier, softer and less expressive tones made HS interpretations stronger). Overall, in their most recent contributions, Niebuhr and Neitsch (2022) have asserted that prosody affects how we perceive and rate HS, and also show how prosodic contrasts make us notice the severity of the message and type of discourse being produced.

In automated detection, we also begin to see works that relate acoustic contrasts to the detection of HS. Boishakhi, Shill, and Alam (2021) begun by considering the properties of the audio signal, as frequency and time domain features, in order to find more precise ways to target HS in a multimodal setting. This study reported satisfactory metrics, achieving high levels of accuracy and recall by adding specific information of all modalities (i.e. audio, image and text). Although they are not directly focusing on prosodic features, they show that extracting features from the audio apart from the textual information benefits the distinction of HS. Lastly, recent works as the one by Bhesra, Shukla, and Agarwal (2024) explore the effectiveness of using prosody-related features in HS detection. They report that the use of audio shows to be more effective than text, and acknowledges that this approach remains relatively unexplored and should be further studied.

In summary, although audio has been used as a modality for detecting HS, it has only been used to detect emotions or extract textual information. Recent approaches have attempted to use the properties of the audio signal to tackle HS, which appears to have some acoustic differentiators. With this work, we intend to contribute to the use of the audio signal and the properties of speech in automatic detection, in order to bring improvements to the multimodal detection of HS.

[2]Features detailed in Section 3.2.2

3 Methodology

3.1 Materials

3.1.1 Collection

In this study, we focus on analysing multimedia that may contain hateful content. We have compiled a database of videos in Spanish, among which figure examples of HS. We focused on collecting videos from the YouTube platform, as here we can find diverse content that can be easily accessed by the general public.

The videos gathered for this task concern themes of gender and equality. The majority of videos have a "YouTube Short" format, videos no longer than one minute and that deal with fixed topics per the short format. These were retrieved manually by making use of specific keywords. The search terms were defined prior to collecting the samples, some were taken from the Spanish subset of the EXIST dataset, a corpus with more than 200 expressions found in sexist contexts (*EXIST*, 2022). This corpus was later enriched with more terms, which allowed us to sample more relevant content [3]. The final corpus of search terms can be seen in Annex A.

Videos were collected in two stages, the first one for our exploratory analysis (see 1 in Annex B), and the second one to extend the first version of the dataset (see 2 and 3 in Annex B). During the first stage, we gathered 100 videos (50 "hate speech" and 50 "non-hate speech"). In the dataset extension, we added 453 videos (203 "hate speech" and 250 "non-hate speech").

All samples were collected by creating Youtube playlists, as we could access them through external APIs and extract information. Videos were extracted by using the Pytube API [4], a library oriented to extracting data from YouTube. All metadata from the videos (e.g. title, length, views...) was stored in descriptive dataframes [5], such that we could have structured and detailed information of the videos in our dataset. Videos that failed to be extracted were automatically discarded. The final dataset consists of a total of 541 videos; 248 "hate speech" and 293 "non-hate speech".

[3] All terms have been defined within the permitted search terms of the platform, as Youtube does not allow to browse videos under offensive prompts, e.g. #feminazi or #feministatóxica.

[4] `https://pytube.io/en/latest/`

[5] `https://github.com/alejaduque/hatespeech_prosody_detection/tree/main/data/datasets`

3.1.2 Annotation

The data were annotated once the samples were collected, where different annotators categorized some of the samples in either of the fixed categories, "hate" and "non-hate". To label the data, we required a definition of HS to ground the concept and decide when a sample fell under this definition. However, defining "hate speech" is complex, as the nature of the concept is highly subjective and there is no formal definition that does not lead to contradictions. Normally, HS descriptions conceptualize it as harassing messages that target specific social groups. The problem with this approach is that it can be interpreted differently depending on the perspective, which makes its definition highly controversial (Kocoń et al., 2021). Moreover, these sort of definitions can be rather vague when it comes to annotation, as it does not encompass the different manifestations of HS can and its complexity.

In this dissertation, we aim to identify HS in discourses that express hatred directly and implicitly in discourse. For this reason, given the potential ambiguity of this concept, we are interested in approaches such as the one introduced by Papcunová et al. (2021), where a series of indicators are defined to objectively diagnose HS. Not only these indicators consider signs, e.g. emotional speech or direct attacks, but they also observe indicators that contribute to implicit HS, such as ambiguous statements, irony or sarcasm. The origin of these indicators try to account for the different perspectives that might arise under the HS concept. We adopted said indicators in order to ground the concept we would use to decide whether a discourse belonged to "hate" class. Additionally, these indicators were enriched with new features that considered cases of reported speech (Chiril et al., 2020). Here we acknowledge that HS can manifest through the invalidation of reported speech and, similarly, that reporting a message does not equate to conveying hatred. The labelling criteria are shown in Annex C.

To label our samples, we took the most representative samples in our dataset and assembled a survey where annotators could read the criteria, visualize and classify the video as they considered (Duque, 2024b) [6]. Not all the videos of the dataset were included in this survey, only the samples that were representative of the full dataset or were ambiguous to label. The survey was answered by four annotators, only two of them reporting prior experience in annotation. All annotators were

[6] See `https://forms.gle/vnU2tKVLCXPTwZTa9`

asked to follow a simple task, which consisted of reading the defined labelling criteria and then proceed to watch each of the enlisted videos (i.e. 10 videos). After visualization, they were able to decide whether the video was hateful, according to their perspective and the indicators given. The final labelling decision was based on whether a given video was perceived as hateful by more than 50% of annotators. Ultimately, the annotation decision was transferred to the dataset. We annotated the dataset based on the similarity of the samples with the examples of the survey. For instance, if a sample in the survey consists of an informal interview where interviewers intend to question or invalidate feminist ideas, and the majority of annotators labelled this content as HS, the rest of the videos with similar formats and traits would be labelled as "hate". Under these foundations, we labelled all of the data to represent our two classes, "hate" and "non-hate".

3.2 Preprocessing of data

In this dissertation we are interested in using the audio modality from multimedia content. More specifically, we want to use the speech data from the soundwaves to observe their prosodic properties. For this, we extracted the audio data from all video samples, allowing us to work with a single modality. The preprocessing stage involved a series of steps aimed to normalize, clean and isolate our data to a more suitable format[7].

3.2.1 Data normalization and adaptation

The first step for preparing our data was to normalize samples in length. Although the majority of samples were no longer than one minute (as mentioned in 3.1.1), we kept some longer videos, as they were considered to have relevant content. Keeping longer data within our samples can be problematic when it comes to processing, as we could find some samples might be heavier to process compared to others. To optimize future processing and maintain uniformity, we set a maximum length of two minutes for all samples. Previous to this, we checked that longer samples contained relevant content within the two-minute time-frame, i.e. the content of the normalized samples are consistent with the assigned labels. Longer samples that did not match the assigned label were discarded. We used the MoviePy library (Zulko, 2017) to shorten all samples to the

[7]The code for preprocessing is available in"pre_processing_audios.ipynb" in project's repository (Duque, 2024a).

maximum length in case they surpassed such limit.

With the normalized samples, the next step was to prepare audio data to isolate the speech from any noisy background. Considering that our data comes from a social network mainly focused on entertainment, it was expected that the audios in our dataset included background elements as music or ambient noise. To isolate the speech, we prepared a program to clean our data by eliminating any unwanted background sound. For this we used the Spleeter tool (Hennequin, Khlif, Voituret, & Moussallam, 2020), an audio-processing library oriented to perform segmentation. This tool uses pre-trained models based on Convolutional neural networks (CNN) to separate different elements in music. Through this tool, we were able to segment our audio-data in two channels, vocals and accompaniment. Out of these outputs, we kept all samples from the vocals channel, where the speech was preserved without any background element.

For all samples, the final outputs obtained were the soundwaves with a maximum length of two-minutes and only containing the speech of the interlocutors. These data were used to perform feature extraction and perform further analyses.

3.2.2 Feature extraction

This step consisted in using the preprocessed data to extract its features and get relevant insights about the prosodic characteristics of speech. For feature extraction, we used of the OpenSMILE toolkit (Eyben, Weninger, Gross, & Schuller, 2013), a library used for audiovisual processing. This toolkit allows to make feature extraction at different levels, as it can extract summarized or frame-wise features from audio data.

Given the different purposes of our methodological process, we have chosen to make use of the summarized and frame-wise information at different parts of the task. In the exploratory analysis, we used OpenSMILE's **functional** features, which corresponds to summaries of the statistical metrics of the features. For the implementation of our Support Vector Machine (SVM) classifier, we make use of the **low-level** features, information extracted in 50 ms chunks of the soundwave, which allows to see the evolution of the features in time.

We extracted features of speech that were informative about intonation, stress and rhythm. How these features are retrieved are provided by Eyben (2015). Firstly, we selected features that reflected acoustic variation in speech:

- F_0 **(Pitch)** : fundamental frequency of the speech signal. Extracted on a semitone frequency scale, starting at 27.5 Hz. This feature is closely related to pitch, as F_0 refers to the physical signal produced by the vocal tract, and *Pitch* refers to how we perceive said signal. It refers to the average number of oscillations per second and expressed in Hertz (Hz). In prosody, as this feature is typically non-stationary and changes constantly, can be used to observe how the pitch changes for expressive purposes (e.g. emphasis, questions, anger) (Bäckström et al., 2022).

  F_0 values were normalized, given that males tend to have lower voices than females and children, which could affect how data is interpreted. We normalized it by using the scaler provided by the scikit-learn library (Pedregosa et al., 2011), which removes the mean and scales values to a unit variance:

  $$z = \frac{x - \mu}{\sigma},$$

  where x is the value we want to normalize, μ is the mean of the samples and σ the value for the standard deviation.

- **Loudness**: the energy and perceived signal intensity from an auditory spectrum. In speech processing, the energy is commonly computed by using the Root Mean Square (RMS) energy, defined as:

  $$RMS = \sqrt{I},$$

  $$I = \frac{1}{N} \sum_{n=0}^{N-1} w^2(n) x^2(n),$$

  where, $w(n)$ is the window function, x is the signal and N number of samples. However, since this metric does not take into account the properties of the human perception of loudness, OpenSMILE uses an approximation of the loudness. According to Eyben (2015) the metric is obtained as follows:

  $$E_i = \left(\frac{I}{I_0}\right)^{0.3},$$

  where I is the signal intensity defined as the signal energy E of $x(n)$, where $x(n)$ has been weighted with a Hamming window function.

We also extracted features related to micro-prosodic variations in short-time frames:

- **Jitter**: the variation of the length of F_0 from one period to the next. Here, an absolute value of the local jitter is given and extracted the average jitter per time-frame:

$$\overline{J}_{pp} = \frac{1}{N' - 1} \sum_{n'=2}^{N'} |T_0(n') - T_0(n' - 1)|.$$

  N' represents the number of pitch periods, the length of a period $n' - 1$ is $T_0(n' - 1)$, and the length of the second period is T_0.

- **Shimmer**: the amplitude variation of consecutive voice signal periods. As with the jitter, this tool uses an absolute metric to measure shimmer. The period to period amplitude is expressed as:

$$\overline{S}_{pp}(n') = |A(n') - A(n' - 1)|,$$

  with peak to peak amplitude:

$$A(n') = x_{max,n'} - x_{min,n'}.$$

- **Harmonic-to-Noise Ratio (HNR)**: the ratio of the energy of harmonic signal to the energy of the noise (i.e. signal components). The noise energy is computed by subtracting the average waveform from each individual waveform and computing the RMS energy E_{noise} of the remaining signal over all N periods. The harmonic energy E_{harm} is computed as RMS energy of the average waveform (Eyben, 2015). The ratio is obtained as:

$$HNR_{wf} = \frac{E_{harm}}{E_{noise}}.$$

Finally, we extracted information from an spectral descriptor:

- **Hammarberg Index**: spectral information related to vocal effort. This metric is often used in the detection of emotional speech, as it used as an energy distribution measure averaged across the utterance (Schmidt, Janse, & Scharenborg, 2016). Defined as a static pivot point where the low and high frequency regions are separated. This measure is the ratio of the

strongest energy peak in the 0–2 kHz region to that of the strongest peak in the 2–5 kHz region. This index is computed as:

$$\eta = \frac{max_{m=1}^{m_{2k}} X(m)}{max_{m=m_{2k+1}}^{M} X(m)},$$

where $X(m)$ is the spectrogram of the samples and max_{2k} is the highest spectral bin index where $f \leq 2$ kHz is still true.

Once these values were obtained from our data, we used them for evaluating the patterns within the two groups by using the functional and low-level values.

3.3 Exploratory analysis

Our first analytic approach consisted of a preliminary observation to assess how well our data represented the two fixed classes. This inspection was done after the first sampling of the data, to determine the viability of extending the database for further analyses.

The exploration was done with two purposes, analysing the means of the groups and how separable both classes were with the selected prosodic features. These analyses were done by using the **functional** values from OpenSMILE, summaries of statistical information of the sample's features (e.g. arithmetic means).

3.3.1 T-test

The first analysis was done to observe the differences between populations in the first portion of collected data. We performed an independent t-test, an statistical test used to determine whether there is a difference between two independent sample means. Here, this test is only a preliminary observation that will allow us to assess how effectively our data is represented by the two fixed classes. We look to compute the following:

- ***P*-value**: which quantifies the probability of observing as or more extreme values assuming the null hypothesis is true. A p-value larger than a chosen threshold (e.g. 5% or 1%) indicates that our observation is not so unlikely to have occurred by chance. Therefore, we do not reject the null hypothesis of equal population means. If the p-value is smaller than our threshold, then we have evidence against the null hypothesis of equal population means (Kim, 2015).

- ***T*-statistic**: a value that help us define the distribution of sample means in relation to the variance, which is used to know the distribution of the sample means, known as *t*-distribution [8]. This will determine whether to support or reject the null hypothesis, as it will be informative of how far this mean differences are in both populations.

To perform this test, we used the Scipy tool, which automatically performs the test when given the mean values of the two populations (Virtanen et al., 2020). We performed a t-test for every feature extracted, namely those described in 3.2.2.

Although we used this test to see the differences between groups, this test was not decisive in which features to select or how data were analysed. This is because the means of the groups did not provide any insights about the distributional patterns of the acoustic features within the two groups, i.e. how they evolve in time and its patterns. Instead, it compresses information in summarized single values. As we are interested in finding patterns of how prosody interacts within HS, we are interested in the trajectory of these features and how it can give us cues of the presence of some of its indicators (e.g. irony or emotional language). This test gave us a first insight into how the two groups differed in general terms but, in further analysis, we would be needing more fine-grained data that allows us to extract patterns.

3.3.2 Support Vectors Machine (SVM)

In the second analysis, we made use of a pattern recognition algorithm to see the separability of our data. Our goal was to determine whether the information we extracted was enough for an algorithm to extract patterns and separate our samples into groups. We use a Support Vector Machine (SVM), a popular algorithm used for classification, regression and detection. SVM is a decision machine, which intends to draw a decision boundary based on the concept of margin, defined as the smallest distance between decision boundary and any of the samples (Bishop & Nasrabadi, 2006).

In Figure 1, from Bishop and Nasrabadi (2006)'s book, we see how the margin defines the distance between the decision boundary and the closest data points. By maximizing said margin, the decision boundary is drawn based on the closest data points, the support vectors. This kind of model intends to find the optimal hyperplane that minimizes the probability of error.

[8]Note that, the two samples would display a normal distribution, as they would have an equal variance because they were independently extracted from an identical population that have a normal distribution (Kim, 2015).

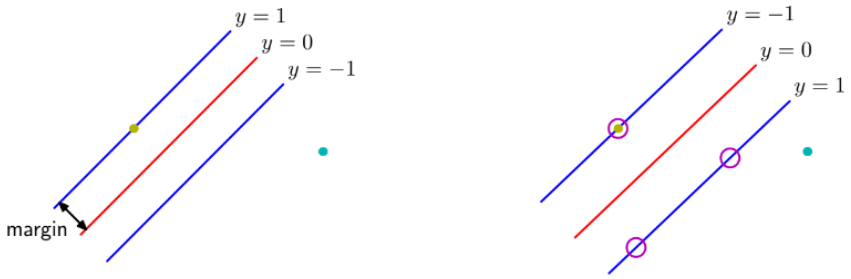

Figure 1: Example of the margin (left) and drawn decision boundary (right).

Under this concept, we want to use this algorithm to determine if we can overfit our data, namely obtain a decision boundary that separates our data into groups as closely as possible. By overfitting our data, we will observe whether the information we have collected in the feature extraction phase is sufficient to separate the two groups, "hate" and "non-hate".

We used the scikit-learn library for implementing this model (Pedregosa et al., 2011). We overfit our data by training a supervised learning model with all the samples. The model is trained using vectors—one per sample—containing the functional values of the features selected and, separately, provided the labels for each sample.

As we maintained the default SVM algorithm from scikit-learn, this model uses a Radius Basis Function (RBF) kernel. However, we modified the values of C, the classical penalty term weight of the SVM, and gamma, a parameter of the kernel that decides the extense of the influence of the training examples in the decision boundary (Learn, 2019). The hyperparameters, values of parameters, were selected by performing a Grid search cross-validation, a method that allows to test different combinations of the hyperparameters to find the most optimal values. This way we aimed to obtain the values of the parameters that would return the better precision value, i.e. the values that lead to more correct detection in both classes.

Under this context, we see how using a pattern detection algorithm is useful to have a preliminary view on the separability of our first sampling. This would mean that even though we do not have enough samples to train a model, we are able to detect patterns within its features and differentiate between the two populations. Separating these groups by using this model would also suggest that extracting the selected features from the speech signal is enough to find patterns that our algorithm can use to decide where a sample belongs.

Looking at the separability of classes with these summaries gives us a clue as to how some patterns can be extracted when all features are combined in vectors for representing each data point. This pattern extraction method opens the door into looking at the feasibility of applying this same algorithm to higher-dimensional data, such as low-level descriptors, which can provide more detailed information about the trajectory of the prosodic features.

3.4 Classifiers

Our second approach tested the effectiveness of different classification methods with our data. Our goal was to observe the feasibility of classifying HS by using prosodic features from the audio.

We tried two methods, one traditional machine learning model and a neural network. Both were implemented as supervised methods, where the output of the training data was already known by the model, as we provide the labels.

We made use of the full dataset of 541 samples (described in Section 3.1.1). These samples were divided in train, test and validation subsets, in order to train our model and evaluate how well it generalized. The data were adapted prior to training, as each model required input formats. The following sections are dedicated to describe the implementation of the models.

3.4.1 Support Vector Classifier (SVC)

As a first test, we used the traditional machine learning method previously applied in the exploratory analysis, SVM. As we have described in 3.3.2, this model is a decision machine that uses support vectors to find the smallest distance between samples and draw boundaries. We use this approach for two purposes, to see the separability of the classes with the full dataset and to observe its effectiveness in classification, thus testing it as a Support Vector Classifier (SVC).

Differently to the exploratory analysis, for this application we will be using the **low-level descriptors** from OpenSMILE. These descriptors provide segmented information about the evolution of each feature over time. For this task we prefer using descriptors rather than functional scores, as this allows the model to identify patterns within each group based on the temporal representation of prosodic properties. Therefore, in this test, we extracted the features selected in 3.2.2 as low-level descriptors.

Prior to training, the dimensions of the resulting matrices from the feature extraction were adapted; given that some videos vary in length, causing the dimensions to differ. We introduced zero-padding, a technique that allowed to equalise the dimensions by adding zeros to shorter matrices to match the size of the largest matrix. This improved data processing by preventing errors that could arise from differences in dimensions.

We first examined the separability of the classes by overfitting the training split. We replicated the test from our exploratory analysis (Section 3.3), using cross-validation to identify the optimal hyperparameters for class separation. As this test showed satisfactory results, this prompted the application of a second test with an SVC.

The classifier was trained using the train, test, and validation subsets, also employing cross-validation to find the best hyperparameters. The key difference in this phase was that we reserved the test set for the final assessment, in order to evaluate the performance of the SVC. The optimal hyperparameters are displayed in Figure 1.

	C	**Gamma**
Overfitting test	1.83	2.34×10^{-6}
Classifying test	20.7	2.22×10^{-7}

Table 1: Optimal hyper parameters for Support Vector Machine (SVM) approaches.

We focused on finding the best values for C and gamma that returned the best decision boundary. As we mentioned in 3.3.2, C is a penalty term and gamma a parameter of the kernel.

Overall, the results of this initial test showed favourable results with the data we fed into the model, prompting us to test more powerful approaches and see if these could analyse our samples similarly.

3.4.2 Convolutional neural networks (CNN)

The second approach taken was to use a neural network to classify. We used a Convolutional neural network (CNN), a method commonly applied for analysing and classifying image data. For our purpose, given that an audio can be interpreted as an image, we used this to observe to what extent we can classify with audio only.

We made use of the preprocessed audio signal obtained in Section 3.2 to generate spectrograms. These are two-dimensional representations of the speech signals, generated by using a Short-Term Fourier Transform (STFT) and extracting the discriminative features automatically (Jahangir et al., 2021). The result is a two-dimensional representation where the x-axis represents time, and the y-axis the energy of amplitude at an specific time (represented in different colors).

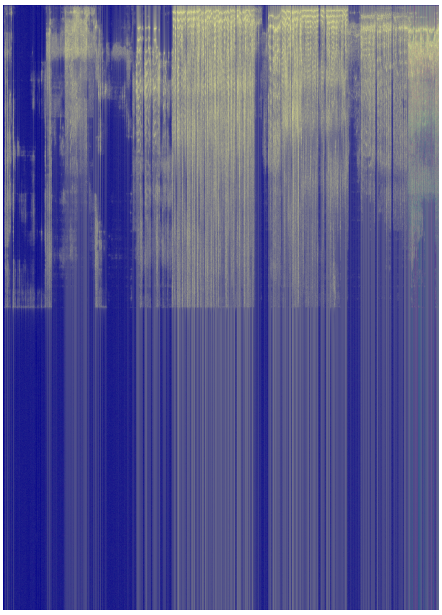

Figure 2: Example of a spectrogram obtained from our data

Figure 2 shows an example of how our samples can be interpreted as images, i.e. spectrograms. These spectrograms where generated by fixing a window length of 2^{12} samples per segment, so the spectrogram is as "squared" as possible and there is a balance between the resolutions of frequency and time. Additionally, we use a power of 2 in the window's length to take advantage of the Fast Fourier transform. Finally, we use a rectangular window, as it is the window that provides better frequency resolution. These parameters are set in order to get the best resolution and balance between the frequential and temporal resolution. With the generated spectrograms, a CNN model was assembled in order to analyse our dataset.

We relied on the architecture of ResNet152 for building the CNN, a model designed to learn from residual functions of previous layers (He, Zhang, Ren, & Sun, 2016). To form the network, these layers are stacked such that residual blocks are on top of each other. The network contains 152 layers in its architecture. Furthermore, to adapt the model for classification we extended the

output by adding two layers of multi-layer perceptrons. The first for introducing non-linearity by adding an hyperbolic tangent function, and the second to include a sigmoid function that would place our result in a value between 0 and 1.

The model is trained by using binary cross-entropy loss function and setting a learning rate of 10^{-5}. To find the number of iterations that returned the best loss, we trained our model over 100 iterations, to determine where the lowest loss was achieved.

Prior to implementing the model, we first attempted to overfit our full dataset, as a way to see how well the classes could be separated with this method. For this, we did not split our data into subsets, but rather introduced the model with all the data available, to see whether any differences could be found. Given that the results of this test were not felicitous, we decided to not attempt classification using this approach, as the classes were not being separated with the given model. Further details will be provided in 4.

4 Results

This section is dedicated to show and discuss the results obtained from the exploratory analysis and classifiers. Most of the following results are displayed in the form of confusion matrices, useful for assessing the final predictions from our models. All methods using SVM and CNN used accuracy scores to asses their performance (Baratloo, Hosseini, Negida, & El Ashal, 2015), this was obtained as follows:

$$\text{Accuracy} = \frac{\text{Nº correct predictions}}{\text{Total predictions}} = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP and TN stand for true positives and negatives; FP and FN stand for false positives and negatives. All positives ("hate") are represented by 1 and negatives ("non-hate") are represented by 0. This score was used to asses how accurately the model predicted both classes. In the following sections we will review the results obtained from these assessments.

4.1 Exploratory analysis

As mentioned in Section 3.3, this exploration of the first sampling of data was done in order to see whether there were differences between the two groups, "hate" and "non-hate". For this, we used

the functional information from the feature extraction, for which we first performed a t-test, and then attempted class separation using SVM.

The t-test informed us of the differences of the means between the groups. Table 2 displays the results of the independent t-test. What these values indicate is that, out of the means of the selected features, there are only two features that have significant differences in their means. F_0 and Jitter seem to differ in their mean of both populations and reject the null hyphotesis. For the rest of features, the groups seem to reflect there are no differences in their means. However, as previously mentioned, given that the mean values do not provide any information about the trajectory of the features, this test was not decisive in what features we would use for our analysis. This test rather shows that there are differences in some of the features, which indicates there are some differences between both classes.

	P-value	T-statistic
F_0	**0.04**	**-2.03**
Loudness	0.79	-0.26
Jitter	**0.003**	**-3.03**
Shimmer	0.26	-1.12
Hammarberg index	0.45	0.75
HNR	0.35	0.93

Table 2: T-test results from exploratory analysis.

For our second test, as explained previously, we overfitted our data using vectors with the means of all features for every sample in our dataset. The results are shown in Figure 3.

We see these results show that most of our data were successfully separated and accordingly classified. Using the available functional information about these prosodic features appears to be enough for successfully differentiating the two populations. Combining all features into vectors allowed the algorithm to find patterns and separate both classes almost entirely. This prompted us to test more powerful methods with more data, this time to test the feasibility of classifying by using more powerful approaches.

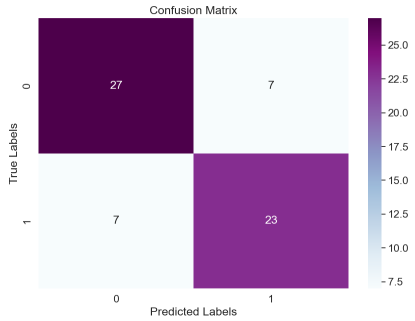

Figure 3: Confusion matrix from SVM overfitting in exploratory analysis.

4.2 Classifiers

In the implementation of the classifiers we used two methods, SVC and CNN. First, we tested how well they separated the different versions of the data, and, lastly, we assessed their performance as classifiers.

The results obtained in the overfitting phase using SVC are displayed in Figure 4, which comprises the use of low-level descriptors for overfitting the samples in our training set.

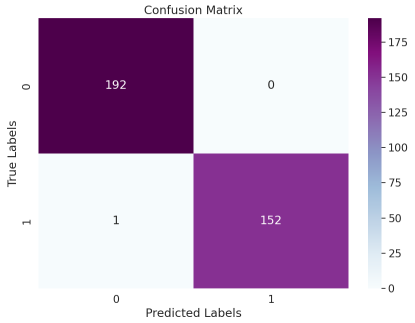

Figure 4: Overfitting of full dataset with SVM.

Results in Figure 4 show that classes were successfully separated, scoring a 99% of accuracy. This means our model was able to successfully find patterns from frame-wise information and draw a decision boundary that separated classes in our training set almost entirely. Given these results, once we verified that classes could be separated with this method, we tried using this model and data for classification. Results from predicting the test set [9] are shown in Figure 5 .

[9] 109 samples

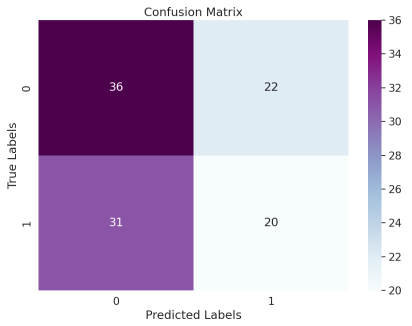

Figure 5: Confusion matrix of SVC.

Results in Figure 5 show a 51% of accuracy from the model, meaning that more than half of the data were correctly detected and the remaining 49% of samples consist of misdetections. We see how despite the success the in overfitting phase, the classification method does not seem as effective in accurately predicting all cases. This suggests the model could be improved for future applications.

At last, for the CNN model implemented using spectrograms, we tried overfitting with the model using the full dataset. We did this in order to see whether the classes were separable given this model and versions of the data. Figure 6 shows results from this test.

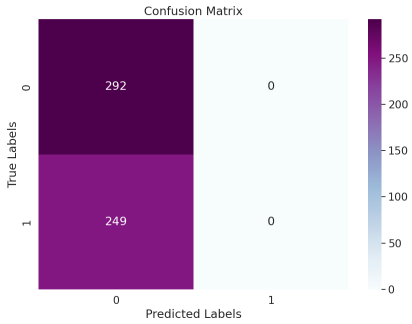

Figure 6: Overfitting results with CNN

The predictions obtained using the CNN showed rather unfortunate results. Figure 6 shows that the neural network is consistently predicting zeros for all the samples presented. With this approach, we obtained an accuracy score of 50%, as the only data correctly categorized are the ones from class 0, “non-hate”. This means our model is not being properly trained on the given dataset, which cause could stem either from model or the data used in training.

As our first attempt in overfitting resulted in such errors, we decided to set aside the implementation of the classifier, as our current proposal was not able to distinguish between classes. Therefore, as getting this model to work requires closer examination and more time, this proposal must be fully revised in future works.

4.3 Discussion

The results from the tested models returned mixed outcomes, as only one of them gave favourable results for classifying HS. Nevertheless, we consider that the results obtained in these tests were rather positive, as we demonstrated that classification of HS is possible when relying on prosodic information.

In SVC, we had exclusively made use of information that concerned prosody and the trajectory of its features. Our results pointed that using these descriptors is sufficient for separating the classes. However, even though the classification was not as effective as the overfitting, these results seem to point that the parameters of the model could potentially be modified to improve HS detection based on prosodic information.

Contrarily, the CNN application was not felicitous despite being the model with higher complexity. Even though the training did not exclusively use prosodic information, but rather the spectrogram, we expected that the prosodic differences would reflect in the images and, consequently, make classes as separable as with the low-level descriptors. Results suggest that there might be more intricate reasons behind the poor performance of the CNN, namely how it seemed to not be learning from data.

The high complexity of the model makes it non-trivial to locate improvement points. We consider its failure could stem from different factors. It could stem from the model's setting, suggesting that the architecture and parameters used might not be appropriate for processing this data. However, it could also stem from the data, suggesting that the parameters for generating the spectrograms might not be returning good enough representations for differentiating the classes.

Overall, both applications are candidates for future revision and improvements. Nevertheless, the results from SVC seem to demonstrate that prosodic information can result relevant in detection of HS, thus demonstrating that prosodic information can be used in HS detection.

5 Conclusions and Future Work

To conclude this work, we see that it is possible to make use of prosodic features to target HS. We have found that training a SVC using frame-wise information of prosodic features makes it possible to separate the "hate" and "non-hate" classes. Additionally, we propose a SVC classifier that in the future can be improved to use this data for HS classification. Therefore, resulting on a positive pathway towards the use of these properties as a modality within multimodal HS detection.

As we initially hypothesized, we can make use of the prosodic information to distinguish HS samples with machine learning approaches. This could suggest that in multimodal environments, where other modalities might not be conclusive, prosodic information from audio could potentially complement these modalities and solve possible ambiguities.

Given the results obtained, future lines of work could be oriented towards optimizing the proposed methods, or explore different approaches that include prosody in the detection of HS.

Considering the outcomes of our CNN, another focus could be improving the quality of the spectrograms in combination to the structure of the CNN model. Aiming to obtain a functional model capable of discriminating between the two classes from images. For instance, new applications could involve the use of different architectures used in CNNs, such as U-net, as a way to test if the use of a different architecture improves training and class separability.

Alternatively, other possible lines of work could involve testing the effectiveness of our proposal in multimodal tasks. That way determine whether using prosodic could potentially work as an additional modality and disambiguator in multimodal HS detection tasks. This opens the door for future works to test if this is true, and test if enriching the sources of information makes a good path to effectively target toxic speech.

6 Appendices

A Key search terms

1. #8m
2. #abortolegal

3. #abusossexuales
4. #cambiodesexo
5. #desigualdad
6. #diadelamujer
7. #díadelamujer
8. #feminazi
9. #femismo
10. #igualdaddegenero
11. #machismo
12. #maternidad
13. #micromachismos
14. #patriarcado
15. #vuelvealacocina
16. La mayoría de las mujeres
17. a las mujeres hay que
18. acoso callejero
19. acoso sexual
20. acoso sexual
21. baja por regla
22. barbie
23. brecha salarial

24. cosificación
25. cultura de la violación
26. cómo una chica
27. despatarre masculino
28. doble discriminación
29. el daño que el feminismo
30. el daño que las mujeres
31. empoderamiento
32. entender a las mujeres
33. estereotipos de género
34. f3min4z1
35. feminazi
36. feminismo
37. feminista
38. feminista rádical
39. gorda
40. hembrismo
41. hombre proveedor
42. igualdad de género
43. las mujeres trans no son mujeres
44. las mujeres trans son mujeres

45. lenguaje inclusivo
46. ley trans
47. macho opresor
48. matriarcado
49. micromachismo
50. misoginia
51. mojigata
52. mujer al volante
53. no binario
54. no todos los hombres
55. nobinario
56. opresión
57. patriarcado
58. privilegio hombre
59. progre
60. progres
61. qué es ser mujer
62. roles género
63. transfobia
64. violencia machista
65. woke

B Playlists used in data collection

1. First sampling: `https://bit.ly/HSSpanishlist`
2. Second sampling: Non-hate speech `https://bit.ly/NHSlistSpanish`
3. Second sampling: Hate speech `https://bit.ly/HSlistSpanish`

C Annotation criteria for labelling hate speech

1. The message conveyed incites or promotes hatred, discrimination or exclusion of a specific person, group or community.
2. The message conveyed justifies hateful behaviour.
3. The message conveyed is intended to humiliate a specific person, group or community.
4. The speaker(s) restricts or denies the right to freedom of a specific person, group or community.
5. The speaker(s) denies current issues that directly affect a group or community (e.g. gender pay gap).
6. The speaker(s) makes use of slurs, violent nicknames or negative connotations when referring to a person, group or community (e.g. use of terms as “feminazi”, “maricon” or “puta”). Either directly, to a specific person or group, or indirectly, by referencing or making generalised assumptions.
7. The video title, description or content of the video makes use of problematic hashtags, nicknames or symbols (e.g. #MujeraLaCocina).
8. The speaker(s) intend to attack by using stereotyped assumptions or prejudices. Either directly, to a specific person or group, or indirectly, by referencing or making generalised assumptions.

9. The speaker(s) makes use of mockery or irony to invalid or deny current issues that affect a person, group or community.

10. The speaker(s) makes comparisons that could be offensive to a specific person, group or community.

11. The initial message presented in the discussion is manipulated or purposely misinterpreted in order to convey a different message.

12. The speaker(s) verbally attacks a specific person, group or community (e.g. uses insults or aggression towards another person).

13. The speaker(s) uses non-verbal attacks towards a specific person, group or community (e.g. uses insulting gestures, such as raising middle finger).

14. The speaker(s) reports a (third-person) experience and expresses disapproval, dismisses or invalidates the reported event.

15. Arguments are constructed in the base of mockery, sarcasm and irony.

16. The speaker(s) makes use of dog whistles in discourse, contributing to ambiguity in the message (i.e. an expression or statement that has a secondary meaning intended to be understood only by a particular group of people).

References

Baratloo, A., Hosseini, M., Negida, A., & El Ashal, G. (2015). Part 1: simple definition and calculation of accuracy, sensitivity and specificity.

Bhesra, K., Shukla, S. A., & Agarwal, A. (2024). Audio vs. text: Identify a powerful modality for effective hate speech detection. In *The second tiny papers track at iclr 2024.*

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4) (No. 4). Springer.

Boishakhi, F. T., Shill, P. C., & Alam, M. G. R. (2021). Multi-modal hate speech detection using machine learning. In *2021 ieee international conference on big data (big data)* (pp. 4496–4499).

Bryant, G. A. (2010, October). Prosodic contrasts in ironic speech. *Discourse Processes*, *47*(7), 545–566. Retrieved from http://dx.doi.org/10.1080/01638530903531972 doi: 10.1080/01638530903531972

Bäckström, T., Räsänen, O., Zewoudie, A., Zarazaga, P. P., Koivusalo, L., Das, S., ... Alku, P. (2022). *Introduction to speech processing* (2nd ed.). Retrieved from https://speechprocessingbook.aalto.fi doi: 10.5281/zenodo.6821775

Chhabra, A., & Vishwakarma, D. K. (2023, January). A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, *29*(3), 1203–1230. Retrieved from http://dx.doi.org/10.1007/s00530-023-01051-8 doi: 10.1007/s00530-023-01051-8

Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G., & Coulomb-Gully, M. (2020). An annotated corpus for sexism detection in french tweets. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 1397–1403).

Corrales-Astorgano, M., Escudero-Mancebo, D., & González-Ferreras, C. (2018). Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with down syndrome. *Speech Communication*, *99*, 90–100.

De Moraes, J. A. (2011). From a prosodic point of view: remarks on attitudinal meaning. *Pragmatics and prosody: Illocution, modality, attitude, information patterning and speech annotation*, 19–37.

Duque, A. (2024a). *Hate Speech prosody detection.* https://github.com/alejaduque/hatespeech_prosody_detection. (Github Repository)

Duque, A. (2024b). *Labelling of Hate speech: Annotator Evaluation Survey.* https://forms.gle/vnU2tKVLCXPTwZTa9.

Ekberg, M., Stavrinos, G., Andin, J., Stenfelt, S., & Dahlström, Ö. (2023). Acoustic features distinguishing emotions in swedish speech. *Journal of Voice*.

Ephratt, M. (2011). Linguistic, paralinguistic and extralinguistic speech and silence. *Journal of*

pragmatics, *43*(9), 2286–2307.

Erekson, J. A. (2010). Prosody and interpretation. *Reading Horizons: A journal of literacy and language arts*, *50*(2), 3.

EXIST. (2022). http://nlp.uned.es/exist2022/.

Eyben, F. (2015). *Real-time speech and music classification by large audio feature space extraction*. Springer.

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st acm international conference on multimedia* (pp. 835–838).

Fujisaki, H. (1997). Prosody, models, and spontaneous speech. In *Computing prosody: Computational models for processing spontaneous speech* (pp. 27–42). Springer.

Gussenhoven, C. (2002). Intonation and interpretation: phonetics and phonology. In *Speech prosody 2002, international conference.*

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 ieee conference on computer vision and pattern recognition (cvpr)* (p. 770-778). doi: 10.1109/CVPR.2016.90

Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of memory and language*, *88*, 70–86.

Henderson, R., & McCready, E. (2018). How dogwhistles work. In *New frontiers in artificial intelligence: Jsai-isai workshops, jurisin, skl, ai-biz, lenls, aaa, scidoca, knexi, tsukuba, tokyo, november 13-15, 2017, revised selected papers 9* (pp. 231–240).

Hennequin, R., Khlif, A., Voituret, F., & Moussallam, M. (2020). Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, *5*(50), 2154. Retrieved from https://doi.org/10.21105/joss.02154 (Deezer Research) doi: 10.21105/joss.02154

Jahangir, R., Teh, Y. W., Nweke, H. F., Mujtaba, G., Al-Garadi, M. A., & Ali, I. (2021). Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. *Expert Systems with Applications*, *171*, 114591. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417421000324 doi: https://doi.org/10.1016/j.eswa.2021.114591

Kim, T. K. (2015). T test as a parametric statistic. *Korean journal of anesthesiology*, *68*(6), 540.

Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T., & Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, *58*(5), 102643.

Lahat, D., Adali, T., & Jutten, C. (2015). Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, *103*(9), 1449–1477.

Learn, S. (2019). 1.4. support vector machines—scikit-learn 0.23. 2 documentation. *Support Vector Machines—scikit-learn 0.23. 2 documentation*.

Luengo, I., Navas, E., Hernáez, I., & Sánchez, J. (2005). Automatic emotion recognition using prosodic parameters. In *Interspeech* (pp. 493–496).

Neitsch, J., & Niebuhr, O. (2019). Types of hate speech in german and their prosodic characteristics. In *1st international seminar on the foundations of speech: Pausing, breathing and voice* (pp. 85–87).

Neitsch, J., & Niebuhr, O. (2020). On the role of prosody in the production and evaluation of german hate speech. In *Proceedings of the 10th international conference on speech prosody, tokyo, japan* (pp. 710–714).

Niebuhr, O. (2022). Prosody in hate speech perception: A step towards understanding the role of implicit prosody. In *The 11th international conference on speech prosody, speech prosody 2022* (pp. 520–524).

Niebuhr, O., & Neitsch, J. (2022, November). The truth below the surface: towards quantifying and understanding the evaluation of german and danish hate speech with eeg biosignals. *Journal of Speech Sciences*, *11*, e022004. Retrieved from `http://dx.doi.org/10.20396/joss.v11i00.16153` doi: 10.20396/joss.v11i00.16153

Papcunová, J., Martončik, M., Fedáková, D., Kentoš, M., Bozogáňová, M., Srba, I., . . . Adamkovič, M. (2021, October). Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex amp; Intelligent Systems*, *9*(3), 2827–2842. Retrieved from `http://dx.doi.org/10.1007/s40747-021-00561-0` doi: 10.1007/s40747-021-00561-0

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E.

(2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Reyes, A., & Saldívar, R. (2022). Linguistic-based approach for recognizing implicit language in hate speech: Exploratory insights. *Computación y Sistemas*, *26*(1), 101–111.

Robert, J., Webbie, M., et al. (2018). *Pydub.* GitHub. Retrieved from `http://pydub.com/`

Saka, P. (n.d.). Hate speech. In *How to think about meaning* (p. 121–153). Springer Netherlands. Retrieved from `http://dx.doi.org/10.1007/1-4020-5857-8`$_5$ doi: 10.1007/1-4020-5857-8$_5$

Sato, H., Mitsukura, Y., Fukumi, M., & Akamatsu, N. (2001). Emotional speech classification with prosodic prameters by using neural networks. In *The seventh australian and new zealand intelligent information systems conference, 2001* (pp. 395–398).

Schmidt, J., Janse, E., & Scharenborg, O. (2016). Perception of emotion in conversational speech by younger and older listeners. *Frontiers in psychology*, *7*, 184571.

Tomasello, R., Grisoni, L., Boux, I., Sammler, D., & Pulvermüller, F. (2022, February). Instantaneous neural processing of communicative functions conveyed by speech prosody. *Cerebral Cortex*, *32*(21), 4885–4901. Retrieved from `http://dx.doi.org/10.1093/cercor/bhab522` doi: 10.1093/cercor/bhab522

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2

Wagner, M., & Watson, D. G. (2010, May). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, *25*(7–9), 905–945. Retrieved from `http://dx.doi.org/10.1080/01690961003589492` doi: 10.1080/01690961003589492

Wierstorf, H., Wagner, J., Eyben, F., Burkhardt, F., & Schuller, B. W. (2023). audb – sharing and versioning of audio and annotation data in python. *arXiv preprint arXiv:2303.00645*.

Zulko. (2017). *MoviePy 1.0.2 documentation.* `https://zulko.github.io/moviepy/`.