Universidade do Minho
Escola de Engenharia
Departamento de Informática

Paulo Jorge Mendes

Online-SoBA: Text Analysis to study Social Behaviors

June 2022

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Paulo Jorge Mendes

Online-SoBA: Text Analysis to study Social Behaviors

Master dissertation
Integrated Master's in Informatics Engineering

Orientador
Pedro Rangel Henriques
Supervisor
Cristiana Esteves Araújo

June 2022

AUTHOR COPYRIGHTS AND TERMS OF USAGE BY THIRD PARTIES

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Paulo Mendes

Assinado por: Paulo Jorge
Costa da Conceição Mendes
Identificação: BI15171514
Data: 2022-06-29 às 23:24:30

ABSTRACT

This document reports a Master's Project that fits in the 5th year of the Integrated Master's in Informatics Engineering from the Universidade do Minho. This Master work, Online-SoBA (Online Social Behavior Analysis), aims at developing a semi-automatic system capable of analyzing short texts that correspond to comments written in reaction to 'posts' in Portuguese social media (social networks or online newspapers), in order to identify behaviors that characterize social opinions about a given topic along a given period of time. To meet the challenge posed, these short texts are analyzed and the information is extracted according to schemes that describe the phenomena to be studied, that is, sentence structures in natural language that characterize the referred behaviors. In order to achieve this goal, a pipeline was implemented to process the texts contained in the *NetLang Corpus*, and later a web platform was developed to present the generated results to the end user, as well as to allow a fluid navigation over them. The report introduces the system architecture designed to attain the objectives, discusses the implementation decisions and the development steps, and shows the interface created for a fruitful knowledge extraction. Before closing the document with the conclusions, an experiment conducted is described and the results are analyzed to assess the SoBA from the end user perspective.

Keywords: Natural Language Processing, Social Behavior, Text Analysis, Sentiment Analysis, POS Tagging

RESUMO

A Tese de Mestrado aqui relatada enquadra-se no quinto ano do Mestrado Integrado em Engenharia informática, na Universidade do Minho. Este projecto, Online-SoBA (online Social Behavior Analysis), tem como objectivo desenvolver um sistema semi-automático capaz de analisar textos curtos correspondentes a comentários escritos em português como reacção a 'posts' em plataformas de comunicação social (nomeadamente redes sociais ou jornais online), de modo a que seja possível identificar comportamentos que caracterizam a opinião da sociedade sobre um determinado tema ao longo de um dado período de tempo. Para enfrentar o desafio proposto, esses textos curtos (os comentários) são analisados e a informação extraída segundo esquemas que descrevem o fenómeno em causa, ou seja, estruturas de frases em linguagem natural que caracterizam os referidos comportamentos. Para que este objetivo pudesse ser atingido, foi implementada uma pipeline para fazer o processamento dos textos contidos no *Corpus NetLang* e posteriormente foi desenvolvida uma plataforma web para apresentar os resultados gerados ao utilizador final, assim como permitir uma navegação fluida sobre os mesmos. O relatório introduz a arquitetura do sistema concebida para atingir os objetivos, discute as decisões de implementação e as etapas de desenvolvimento, e mostra a interface criada para uma extração proveitosa do conhecimento. Antes de encerrar o documento com as conclusões, é descrita uma experiência conduzida e os resultados são analisados para avaliar o SoBA a partir da perspetiva do utilizador final.

Palavras-Chave: Processamento de Linguagem Natural, Comportamento Social, Análise de Texto, Analise de Sentimentos, POS Tagging

AGRADECIMENTOS

Antes de mais quero agradecer a toda a equipa do Projeto *NetLang* e a todos aqueles que tiveram um papel ativo na validação da ferramenta desenvolvida.

Um agradecimento especial aos meus orientadores, Pedro Rangel Henriques e Cristiana Araújo. Obrigado pela oportunidade e pela vossa dedicação, paciência e perseverança nesta longa jornada. Seria muito mais difícil concluir esta etapa sem vocês.

Por todos os momentos de colaboração e partilha de ideias, quero agradecer à Filipa Pereira, à professora Idalete Dias e em particular ao professor José João Almeida, que mesmo antes de eu ter iniciado este projeto sempre se demonstrou disponível para ajudar no que fosse preciso e por quem eu tenho um carinho especial.

Deixo aqui um agradecimento muito especial à minha mãe, pois é graças a ti que hoje sou quem sou. Apesar de tudo, quero que saibas que nunca me faltou nada. Esta vitória também é tua, pois sempre me mostraste que devo continuar a lutar e dar o meu melhor.

Claro que não podia deixar de agradecer os meus colegas do ilustre *Grupo Objetivo de Trabalho Aplicado*, que ao longo destes anos a nossa relação foi uma de camaradagem e cooperação nas diferentes Unidades Curriculares.

Quero agradecer à Catarina pelos diversos cafés acompanhados de boa conversa, que por vezes foram essenciais no processo de manutenção da minha sanidade mental.

Quero também agradecer aos dois anjos da guarda que me protegem. Celso, obrigado por teres feito parte da minha vida e me teres ensinado que aquilo que realmente importa não é quanto tempo temos, mas sim a intensidade com que vivemos cada momento. Flávia, obrigado por todo o cuidado, atenção e carinho. Sei que poderei sempre contar contigo.

Por fim, quero agradecer à minha carpa, não só por me ter dado a motivação final para eu concluir esta etapa, mas também por me fazer querer ser a melhor versão de mim mesmo. Quero também agradecer ao meu cão, que me ajudou nesta aventura com a sua companhia e boa disposição.

CONTENTS

LIST OF FIGURES

ACRONYMS

C

CMC Computer-Mediated Communication.

J

JALC Judgment Analysis Lexicon Classifier.

N

NLP Natural Language Processing.

P

POS Part of Speech.

S

SUD Socially Unacceptable Discourse.

SVM Support Vector Machine.

1

INTRODUCTION

The Master Thesis here reported was developed within the framework of the international project *NetLang*[1] – *The Language of Cyberbullying: Forms and Mechanisms of Online Prejudice and Discrimination in Annotated Comparable Corpora of Portuguese and English* (PTDC/LLT-LIN/29304/2017) (Henriques et al., 2019).

1.1 MOTIVATION

With the rise of social networks and online communication, sharing one's opinion with friends, family or even complete strangers has never been easier (Stone and Wang, 2019). This increasing online exchange of ideas may sometimes lead to civil discussions which then can devolve into *Socially Unacceptable Discourse* (*SUD*) (Duggan, October 2014), such as hate, discriminatory, offensive or threatening speech.

These kinds of interactions can be quite harmful towards readers, and although some progress has been made with regards to limiting its proliferation (Jourová), it remains a challenge when trying to accurately detect it on a timely manner (Vidgen and Yasseri, 2019; Zhang, 2018), as well as implement effective prevention mechanisms in order to reduce its damaging influence (Ullmann and Tomalin, 2019).

A big part of the difficulty with automating detection and moderation of this kind of speech comes from the existence of many grey areas, where identifying what sort of violation a comment commits becomes troublesome. If the occurrence of spelling errors, *Computer-Mediated Communication* (*CMC*) specific linguistic features, along with the existence of cultural differences between groups of people of different backgrounds are all taken into account, it will be safe to assume that any rule-based system, that simply looks for a certain word or phrase, will quickly become obsolete.

Therefore, before one is able to develop a more functional approach, it is necessary that a thorough multidisciplinary understanding of this type of communication is acquired, in order to be able to pick up perception of when these views originate and track how they evolve across time.

[1] https://sites.google.com/site/projectnetlang/

This project intends to approach that goal by building upon the *corpora* collected by the *NetLang* Project, developed at Universidade do Minho. In order to help researchers in the human and social sciences or psychologists with this matter, the recovered online posts and commentaries that express this kind of language will be processed and analysed, so that an evolutionary model of its characteristics can be extracted.

1.2 OBJECTIVES

This Master's Work has the following objectives:

- Identify and understand the evolution of social opinion about certain topics across time;
- Compare levels of social stigma and prejudice on certain topics between different time frames and social groups in online platforms;
- Develop a semi-automated system capable of analysing short comments present in the *corpora* extracted in the context of the NetLang Project and produce relevant information, in order to satisfy the previous two points;
- Further develop the software to the point of being able to assist the user in the process of identifying the events that triggered a certain change in social opinion.

1.3 RESEARCH APPROACH

The methodology that was followed to prepare this master thesis is composed of the following steps:

- Bibliographic study to deeply understand the state of the art in the area concerned with the detection of socially unacceptable discourse;
- Analysis and processing of the *corpora* extracted by the NetLang Project;
- Develop a strategy and propose a solution;
- Development, testing and adjustment of the presented solution;
- Evaluation and discussion of results.

1.4 RESEARCH HYPOTHESIS

By analysing posts on Social Media and applying Natural Language Processing techniques, is it possible to develop a tool capable of helping researchers to better identify behaviors that define social opinions about a specific topic along a given period.

1.5 DOCUMENT STRUCTURE

This document is divided in seven chapters. This first one focuses on contextualizing the problem and the motivation behind the project, as well as defining objectives and clarifying the methodology for achieving them.

Chapter 2 contains an overview of the state of the art on the automatic detection and moderation of *Socially Unacceptable Discourse* (*SUD*) and on sentiment extraction and analysis.

Chapter 3 consists of a detailed description of the proposed solution for this project.

In Chapter 4 the problem of data preparation is addressed, along with what were the main challenges faced and what compromises were implemented to overcome them.

A detailed description of the implemented web application can be found in Chapter 5.

Chapter 6 describes an experiment that was conducted to assess the SoBA from the end user perspective

Lastly in Chapter 7, the document is closed with a summary of its content, conclusions, and direction for future work.

2

STATE OF ART

This chapter focuses on presenting the techniques used for detecting *Socially Unacceptable Discourse* (*SUD*) in *Computer-Mediated Communication* (*CMC*). Hate speech is defined by the Cambridge_Dictionary as *"public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation"*, while offensive speech can be defined as speech that might contain derogatory words towards a group, with the intent of upsetting or insulting those on the receiving end.

2.1 LEXICONS

Some of the earliest attempts at solving the problem of automating detection and moderation of *SUD* consisted in the use of lexicons. This approach focuses on building a vast collection of words or expressions deemed undesirable, which can then be queried in order to determine if a particular text can be considered abusive (Hatebase).

One problem with this method is that, even when the lexicon contains all the correct variations for every given word, it is still leaving out cases where improper spelling has occurred. These instances could be the product of human error or, in some cases, when a user intends to engage in *SUD* on a platform with this kind of moderation, they can try to bypass it by replacing a few letters on certain words. An expansion of the lexicon or the addition of a rule based system with the aim of covering all these new deviations, is not only impractical but nearly impossible.

Another shortcoming of a system like this is its inability to understand both the context of the words used and their intent, which is often the case when irony or sarcasm is involved. However, this uncertainty can be reduced by mapping words to vectors, with the help of word embedding algorithms, which can then be processed and their similarity with other words and concepts measured, allowing for the detection of the sentiments rooted in a particular text.

2.2 SENTIMENT ANALYSIS

By identifying the trending views, moods and attitudes of certain groups of users, it is possible to better anticipate the behavior and reaction of the general public to certain events, as well as predict with more accuracy how their communication patterns about these themes will evolve over time (Sobkowicz et al., 2012).

When it comes to the problem of automatic extraction of subjectivity and polarity from a text, there are two main approaches. One, with the help of lexicons, looks at the semantic orientation of words or phrases contained in the text, so then it can calculate its orientation as a whole (Taboada et al., 2011). The second approach performs a kind of supervised classification, using classifiers built from labeled excerpts of texts. Usually systems that follow this approach build *Support Vector Machine* (*SVM*) classifiers, trained on a data set using N-grams and sometimes *Part of Speech* (*POS*) labels. These kinds of classifiers can achieve text polarity detection with a fairly high accuracy rate, however, if they are applied to a domain different from the one trained on, their performance will decrease considerably.

Some researchers have already applied these techniques for developing their own predictive models. By extracting the polarity of user messages on Twitter, Wang et al. (2012) was able to develop a system that analysed tweets about the presidential candidates in the 2012 U.S. election. A similar work was made by Heredia et al. (2018), that decided to train a Convolutional Neural Network with the help of annotated lexicons, to be able to detect the comment polarity. Both of these approaches however, do not go into more detail as to justify the presence of the polarities detected, and are not capable of identifying what sentiments are influencing the users' opinion.

2.3 TOPIC DETECTION

Topic modeling deals with identifying which are the themes of context contained in a particular text. When applied to *CMC* it can facilitate with the identification of emerging societal trends and analysis of public reactions to certain events. The process of topic and sentiment classification is usually done separately, by first detecting a topic and later extracting the corresponding sentiment.

By detecting both topic and sentiment simultaneously, the accuracy of the extracted opinions can be improved, due to the fact that the polarity of an expressed sentiment is often dependent on the topic being discussed. There have been some attempts in making this detection simultaneous, however, most of those rely on some form of topic post-processing to be able to correctly determine the polarity of the sentiments contained in a given text (Dermouche et al., 2014).

3

SOBA, PROPOSED APPROACH

The goal of this project is to be a helpful tool for identifying behaviors that define social opinions of a given period. To achieve this, it applies several NLP techniques on Social Media posts and commentaries. These short texts were recovered and provided by the *NetLang* Project, developed at Universidade do Minho.

3.1 SYSTEM ARCHITECTURE

After carefully analysing the contents of the *corpus*, we have reached the conclusion that there is a need for the application of a NLP Preprocessing Pipeline on the raw files contained in the *corpus*. Only after this first step has been fully implemented can we start to develop the mechanisms that will provide insights on Sentiment Analysis.

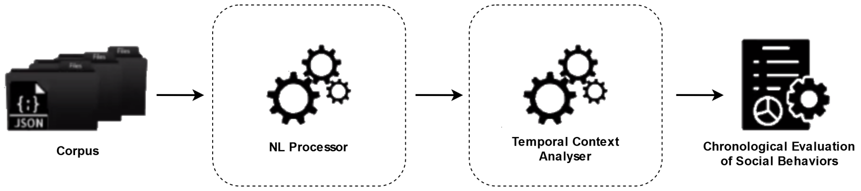

JSON
Corpus
NL Processor
Temporal Context Analyser
Chronological Evaluation of Social Behaviors

Figure 1: System Architecture

This Natural Language Processor (Figure 2) can be broken down into several steps. The first couple focus on text transformation as well as the removal of undesirable elements, such as leftover HTML tags.

Afterwards, it will begin a process of identification and correction of misspelled words. This is being achieved with the use of the *SymSpell* library for the *Python* language. A language dictionary must be provided if this approach is to be successful, and results have been shown to improve by increasing the vastness of the dictionary and by also supplying a more accurate word frequency for each entry. For a few edge cases it was found that, by providing a second dictionary containing a list of likely bi-grams, the outcome of this step could be further refined.

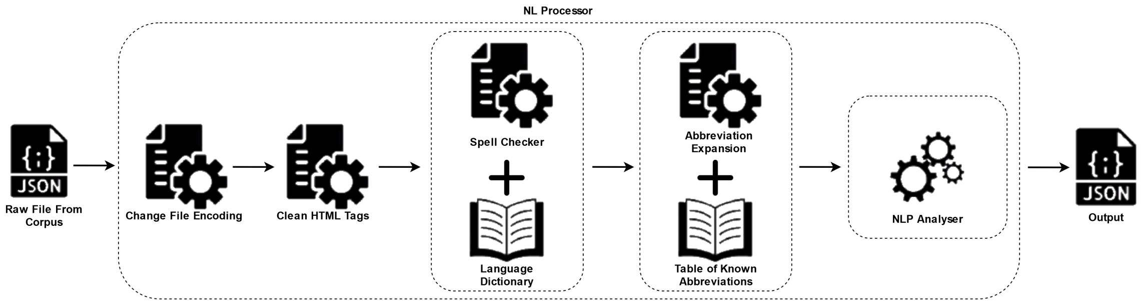


Figure 2: Complete Natural Language Processor Architecture

Following this, the processor will look up a table of known abbreviations and replace with the expanded form any match found in the given *json* file.

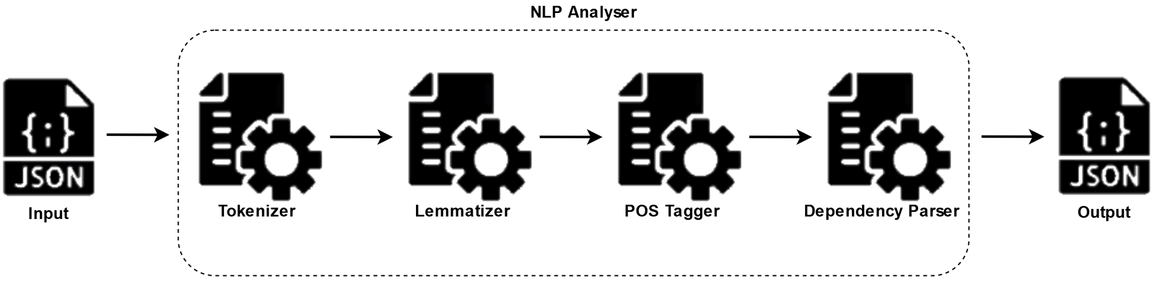


Figure 3: *NLP* Traditional Analyser Pipeline

Finally, after the texts have been cleaned of most of these imperfections, the program will enter the *NLP* Analyser stage (Figure 3). After tokenizing the comments present in the file, it will apply *Part of Speech* tags to each word, while also providing the normalized word equivalent (lemmatization), along with the word's relation regarding the rest of the sentence it is contained in (Dependency Parser).

3.2 DOCUMENT ANALYSIS AND EXPLORATION

Since this first step of implementing the pipeline is a somewhat intricate process, requiring a lot of testing and fine-tuning, a considerable part of the time in this initial project phase was dedicated to the exploration of tools and application of *NLP* techniques, in order to better prepare the data.

After a manual examination of the files contained in the corpus, we got a better grasp of their internal structure and what sort of immediate treatment they required before they were ready to move along the next steps. We soon noticed that some of the texts had leftover HTML tags that would interfere with some of the following steps, such as POS-Tagging. Once this issue was dealt with, the four phases of the NLP Analyser (Tokenization, Lemmatization, POS-Tagging, and Dependency Parsing) were implemented, with the help of the spaCy library[1] for Python. However, after a close inspection of the results produced by this step, we realized Spacy was having trouble tagging misspelled words correctly, which made us search for a way of spellchecking the files since these words consisted of an extensive portion of their respective file.

On the first attempt at solving this problem, the algorithm that was being used employed the Levenshtein distance for selecting the closest correct word off of a list of possible cases when it found a misspelled word (Peter Norvig). After some testing, this method proved much too slow to be used, especially on some of the larger documents. This shortcoming was the necessary motivation to find the current spellchecker (SymSpell), which instead uses the Damerau–Levenshtein distance and has proven to be several orders of magnitude faster than Norvig's algorithm.

After examining the output files, an improvement in the tagger's performance was noticeable, although this time it was struggling whenever an abbreviation of a word occurred. After this initial exploration, the next task consisted in identifying and collecting all the major variants of abbreviations for each word, which would allow their replacement with the expanded equivalent.

[1] https://spacy.io/

4

SOBA, DEVELOPMENT - DATA PREPARATION

This chapter focuses on describing the entire development journey for this project. The techniques and tools used will be presented and the results of their application analysed.

4.1 CORPORA

This project is using the Portuguese side of the *corpora* collected by the *NetLang Project*, developed at Universidade do Minho. *NetLang's* research problem is online hate speech, that is, the expression of prejudice and discrimination on the Internet, involving communicative situations. The *NetLang Corpus* is an annotated comparable *corpus*, that focuses on the English and Portuguese languages.

4.1.1 *File Structure*

The Portuguese segment of the *NetLang Corpus* is composed of around five hundred *json* files, extracted from three sources: two Portuguese online newspapers, *Sol* and *Público*, and from the *YouTube* comment section. These *json* files and their structure were independently designed and generated by the *NetLang Project* team, prior to the start of any of the work reported in this document.

Each file is then organized into two sections: the header and the comment thread. The former contains information regarding the initial post, for instance its title, content, date of posting and how much of an impact it had in terms of views and likes/dislikes. Additionally the header also encompasses data concerning the extraction process, in particular the date of extraction, its source with the original url, as well as few other useful metrics such as the list of identified sociolinguistic variables and their associated keywords.

The latter includes the entirety of the comments, that were present at the time of extraction, displaying for each one it's written content, coupled with additional useful information, namely its creation date and the list of relevant keywords contained within that comment, that were identified during the extraction process.

4.1.2 *Preparation*

Before the files could be used by the later stages of our pipeline, they required treatment. The reason for this was the need to remove some impurities, mainly within the header metadata fields, as well as an overall necessity to normalize values or apply consistent data formats across each entire file.

The type of initial corrections that were made, varied from simply removing an unnecessary word from an integer value and making sure they were being represented exclusively by digits from 0 to 9 (e.g. the values "1000 views", "1 thousand" and "1.000" would all be converted to "1000"). Afterwords, the original post text sometimes, depending on the extraction source, could contain leftover html tags, encoding specific characters, as well as a few specific sentences promoting the online news platform in question or asking for a user login. These cases were quickly corrected with the help of regular expressions.

Eventually the date fields were in need of maintenance. Despite the fact that each one of the three extraction sources had their own unique way of filling the post's publishing date, located in the header of the *json* file, they were all successfully translated into the same format (e.g. the final date of "2020/03/22" would be represented as "Publicado a 22/03/2020", "22 de Março de 2020" and "22 de Março de 2020, 16:20" by YouTube, Sol and Público respectively).

The main predicament presented itself when it was time to normalize the comment dates. While the ones extracted from Público would preserve complete temporal information regarding the day, month and year of publication, the comments extracted from YouTube and Sol would only monitor how long ago the their publication had occurred. At first glance this may not give the impression of being such a massive problem, since by just subtracting this value to the post's extraction date one could attain the actual comment date. However, a more attentive reader may have already realised how doing this could be troublesome down the line.

Suppose, for the sake of argument, that a post was extracted on 2020/05/25 and contains three comments, where their date fields contain *3 days ago*, *2 months ago* and *1 year ago* respectively. By applying the previously described method to the first value, one would obtain 2020/05/22 with absolute confidence. The same cannot be said about the second example. On account of the original value only displaying how many months ago the comment was created, instead of being able to determine the exact day, we would only be able to arrive at a set of possible dates (e.g. in this case, both 2020/03/20 and 2020/03/07 could be classified as having been posted *2 months ago*). Considering that our model is only concerned with the comments' year and month of creation, this hurdle ends up being mitigated.

The primary dilemma lays within comments posted in over a year ago, since the set of possible valid dates would range over months, instead of merely days as in the previous case (e.g. in this case, both 2019/04/28 and 2018/06/02 could be classified as having been posted *1 year ago*). With the aim of solving this issue, the decision was made to assume those instances occurred in that same month, however many years ago. In spite of this not being the best solution, at the time the alternative was to discard these entries completely.

4.2 NL PREPROCESSING PIPELINE

As formerly mentioned in Subsection 3.2, very early in this project's development period, once the files were prepared, an initial implementation of the *NLP* pipeline was carried out, with the help of spaCy.

4.2.1 *spaCy*

spaCy[1] is an open-source software library for advanced natural language processing. It features convolutional neural network models for part-of-speech tagging, dependency parsing, text categorization and named entity recognition, with prebuilt statistical models capable of performing these tasks on multiple languages, including Portuguese.

Being such a powerful tool, the decision to take advantage of spaCy's capabilities was clear, since it would allow us to speed through the implementation process of the first few steps on our pipeline, namely the tokenization, lemmatization, *POS* tagging and dependency parsing.

However, after a closer look at the results, it soon became clear that this first step was not going to be so straightforward as previously thought. Despite the fact that in some cases the last three steps were presenting acceptable results, usually when the comment being analysed by the tool was written with clearer language and did not contain spelling mistakes, for the most part spaCy was having difficulty performing these steps correctly.

In an attempt to improve the generated results, some experimentation was done with the introduction of a spellchecking step in the pipeline. Even though it could be said that some success was achieved with this effort, the initial problem mostly persisted. This occurrence suggested that the presence of misspelled words might not be the sole responsible for the current predicament. Their existence, when combined with the overall presence of poorly structured sentences, which so often is characteristic of this type of informal online discourse, may explain why these models are struggling so much to yield adequate results.

[1] https://spacy.io/

4.2.2 *Spell Checker*

Regarding the Spell Checker, as mentioned earlier in Subsection 3.2, an initial effort was made on this front with the integration of an algorithm, based on Peter Norvig's work, which made use of the Levenshtein distance. Unfortunately, it soon became clear that this implementation was too time consuming. This limitation became even more evident when processing some of the larger YouTube files.

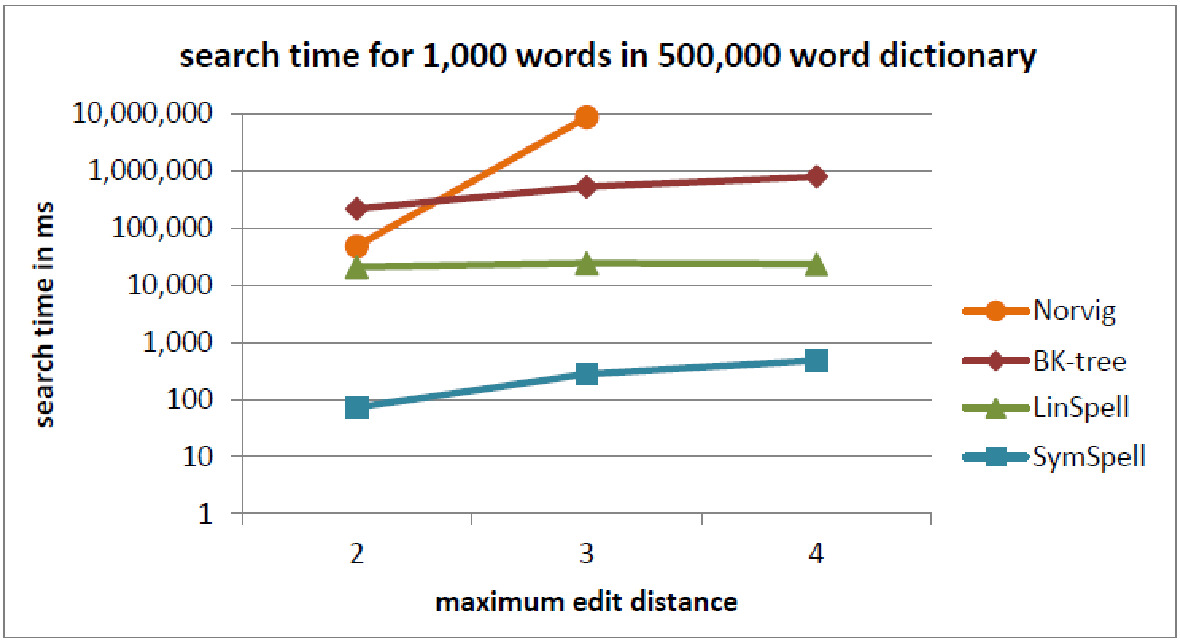


Figure 4: Performance comparison for single term search time between different spelling correction algorithms. Sourced from Wolfgarbe (2021)

The second attempt saw the use of the SymSpell algorithm, which instead employed the Damerau-Levenshtein distance. As it can be observed in Figure 4, this implementation is remarkably faster when compared to the previous attempt.

Once regular words were being corrected in a timely manner, the introduction of a few cases, where word abbreviation occurred, was done to the Portuguese dictionary, generously made available by the *Projecto Natura* team[2], and recomended by its leader, professor José João Almeida, from Departamento de Informática at Universidade do Minho.

[2] accessed at https://natura.di.uminho.pt/wiki/doku.php?id=ProjectoNatura

4.3 sentiment analysis

The initial vision had for this project was for the development of a tool that would be capable of detecting the exact sentiments that could be present within each one of the analysed comments.

For this to be feasible, it was decided to use as a starting point the work developed by the late American psychologist Robert Plutchik on this matter.

4.3.1 *Theory of Emotion*

In his psycho-evolutionary theory of emotions, Robert Plutchik states that there is a small number of basic, primary emotions and that all other emotions are mixed or derived states of varying intensities. In other words, those other emotions occur as a result of a combination of primary ones (Plutchik, 1982) . Furthermore, Plutchik then went on to formalize his theory by organizing these eight prototype emotions (joy, acceptance, fear, surprise, sadness, disgust, anger and anticipation) in a multidimensional model, illustrated by Figure 5, where more closely related emotions are placed within a 90 degree angle of each other.

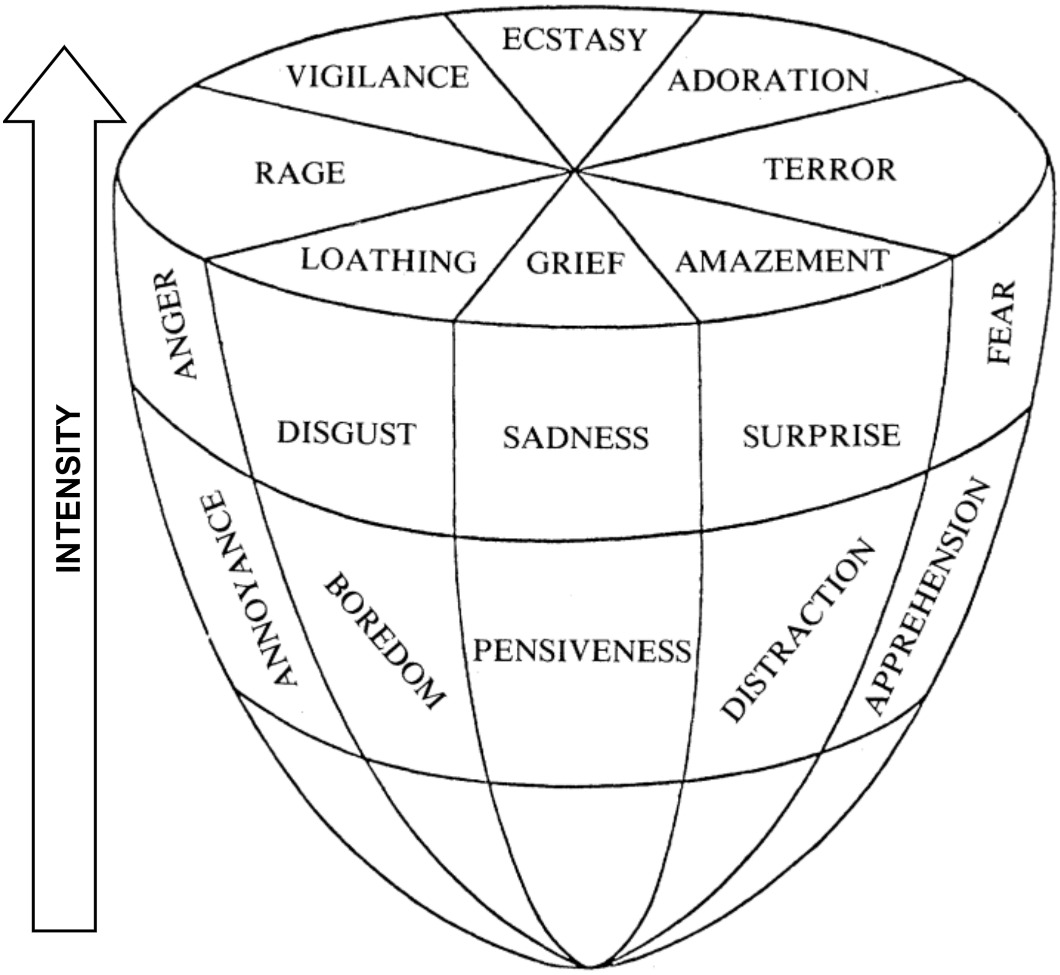


Figure 5: Robert Plutchik's multi-dimensional model of emotions. Adapted from Plutchik (1982)

For each 'slice' of emotion, the vertical dimension implies the presence of a variable of intensity, with the terms at the top representing maximum intensity of each basic emotion dimension. The shape of the model also implies that the emotions become more indistinguishable at lower intensity values.

4.3.2 *Labeled Emotion Dataset*

For the successful application of Plutchik's theoretical model of emotion on the NetLang corpus, a Portuguese dataset containing intensity values for the eight primary emotions was required. The chosen dataset for this task ended up being the XED dataset, developed by the Language Technology Research Group at the University of Helsinki (Helsinki-NLP, 2021). This dataset was created by the initial annotation of the English and Finish versions of movie subtitles, sourced from OPUS (Tiedemann, 2012), and then their translation into 41 additional languages, using annotation projection to aligned subtitles.

This dataset is using multi-label classification and the .tsv file is structured in such a way that the subtitle string being annotated is placed on the first column, while the list of annotated emotions for the respective subtitle is contained by the second column. Regarding the entries on the second column, each emotion is separated by commas and represented by a number between 1 and 8.

However, before the dataset could be used to train the machine learning model, it was transformed with the help of KNIME, by splitting the second column into eight new binary columns. Each one of these columns would indicate if the associated emotion was present in the given subtitle, when the value was equal to 1.

4.3.3 *Machine Learning Approach*

When the dataset preparations were finished, it was time to feed it into a sentiment classification model, built with the help of the Python library Scikit-learn[3]. After its training was complete, the model was applied to a section of the corpus. However, the accuracy of the predictions was not very high.

Unfortunately this was already expected, even before the model was created and applied. During the dataset preparation phase, we came to the conclusion that the quality of this dataset not the best for what we are trying to accomplish, both in terms of the number of unique cases contained in it, as well as the veracity of some of the classifications for many of the entries. Furthermore, the type of informal discourse that characterizes almost all the comments contained in the corpus, significantly diverges from the set of phrases that the authors of the dataset used in its creation.

It is also important to mention that there were some terms found in the dataset where emotions were being identified for no apparent reason. This brought up the assumption that maybe, during the sentiment annotation process, the authors are taking into account the context in which each term occurred. Assuming this is true, it is something that cannot be accessed at the moment.

4.4 POLARITY ANALYSIS

After having seen that the results obtained with the implementation of the previous model were not the most desirable, the initial approach was reviewed and the decision was made to develop a process with a lower degree of complexity. Instead of trying to pinpoint the exact sentiments that are being expressed in a comment, maybe a more successful approach would be to check if a comment contains certain words or expressions, that were previously categorized as being either positive or negative. Afterwords, in order to come up with a

[3] https://scikit-learn.org/stable/

number that classifies the comment as mostly positive or negative, all that was necessary was to look through the set of strings with known polarity that matched that comment, and plug them into a mathematical formula.

4.4.1 *SentiLex-PT*

SentiLex-PT[4] is a sentiment lexicon especially useful for opinion mining applications involving the Portuguese language (Carvalho and Silva). This lexicon contains over eighty thousand entries, consisting of nouns, adjectives, verbs and idiomatic expressions. Each entry has the associated polarity value, varying between positive (1), negative (-1) or neutral (0). These polarities were mostly manually labeled, with some entries being automatically annotated with the help of the *Judgment Analysis Lexicon Classifier* tool, developed by the project team.

With the current implementation, the final polarity value for each comment is being generated by firstly identifying which terms contained within the SentiLex lexicon, adding their respective polarity labels and finally dividing the resulting number by the total amount of terms successfully identified.

This application of the lexicon to this case study yielded highly promising results, in spite of a some cases where the resulting comment polarity was found to be unreliable. These instances usually only occurred when sarcasm, irony or negation could also be found in the same comment.

4.5 THEME DETECTION

Once the comment polarity detection mechanism was outputting satisfactory results, it was time to start the implementation of an independent system, capable of identifying the themes that could be present in each comment. To accomplish this, a decision was made to start with the dictionary approach right away this time. After looking at what was publicly accessible, and not being very pleased with what was available to work with, it was decided that an attempt would be made to develop a custom solution for the missing theme dictionary problem.

Initially, the custom-built dictionary was taking advantage of the fact that the *json* files, from where the comments were being sourced, already had in their metadata information regarding the sociolinguistic variables contained within each file, together with detected keywords that justified their presence. By mapping which keywords were associated with which variable or category, a functional, yet primitive, custom theme dictionary was brought about.

[4] https://b2share.eudat.eu/records/93ab120efdaa4662baec6adee8e7585f

With this primitive version in hand, there was a need to expand it, both in terms of the average number of cases for each category, along with the total number of categories. This expansion was brought about by firstly generating a frequency map of every commented word in the corpus, and then manually going through them, starting with the most frequent, to decide if they could be inserted into a given theme. Of course, this process by itself would not provide a very encouraging outcome, since the most common words in any language are, for what we are trying to fulfill, worthless. This is why a list of Portuguese stop words, so kindly provided by the *nltk* python library, was used to exclude these words from the frequency map. Although this was an essential help, it was still necessary to manually expand the list of dismissed words.

Finally, once the theme dictionary was encompassing enough terms and it was constructing sensible predictions, the enrichment process was halted. Even though at the time of writing this thesis, the current version of the theme dictionary is the one being used in the web interface, it is not to be considered a finished product, for there is still a great deal that can be done to improve upon the work done. Furthermore, owing to the fact that this dictionary's evolution was guided by the author alone, there is a nonzero probability that it might be slightly skewed.

4.6 SUMMARY

In this Chapter it becomes increasingly clear that the task of cleaning and categorizing the data was the most challenging step in the entire implementation process. The development process is being documented with descriptions of the multiple approaches attempted to solve this problem, followed by an analysis of their respective results. As a consequence of the knowledge acquired throughout the Chapter, the initial approach was revised and adjusted until the current working version could be achieved.

5

SOBA, DEVELOPMENT - WEB APPLICATION

Once the tool's performance was adequate, the logical next step was to start working on a web-based user interface, so that researchers could have a more intuitive usage experience. Additionally, this would allow a large portion of the information to be displayed graphically, which could further improve readability.

5.1 INTERFACE

The web interface is organized into three different options for data exploration: *Source Analysis*, *Search by Theme* and *Search by File*.

5.1.1 *Source Analysis*

By selecting the *Source Analysis* option, the user will be taken to a page containing two graphs: the *Sources* graph and the *Contributing Files* bar chart. Both of these are displaying information regarding the whole corpus.

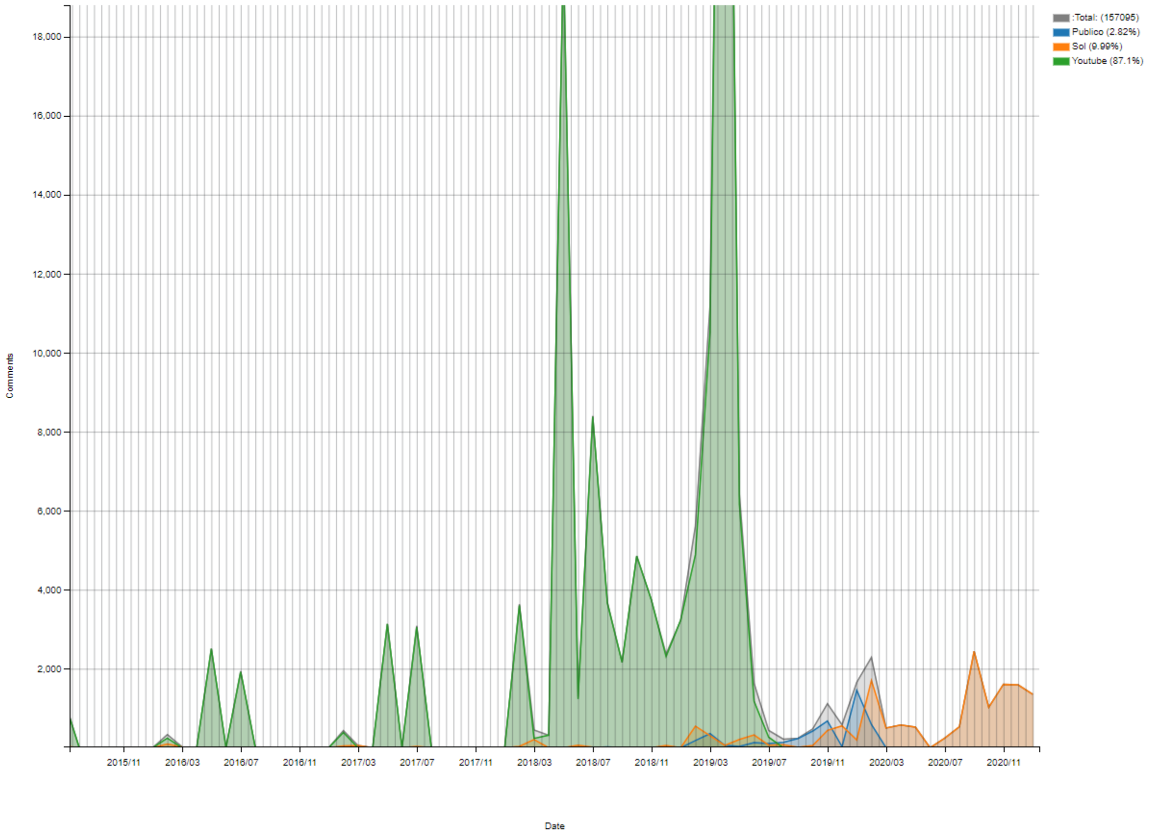


Figure 6: Source graph zoomed in on a time frame between November of 2015 and November of 2020

The first one provides insight about the volume of existing comments for each month, while also identifying the relative contribution that each one of the extraction sources is responsible for, during the entire graph's timeline. Statistics like the exact number of comments in the corpus and the contribution percentage of each source can be found in the legend, as it can be seen in Figure 6. Even though at the time of writing this thesis there are only three different extraction sources being taken into account, with the aim of improving readability, each one of the items in the legend can be toggled *on* and *off* independently.

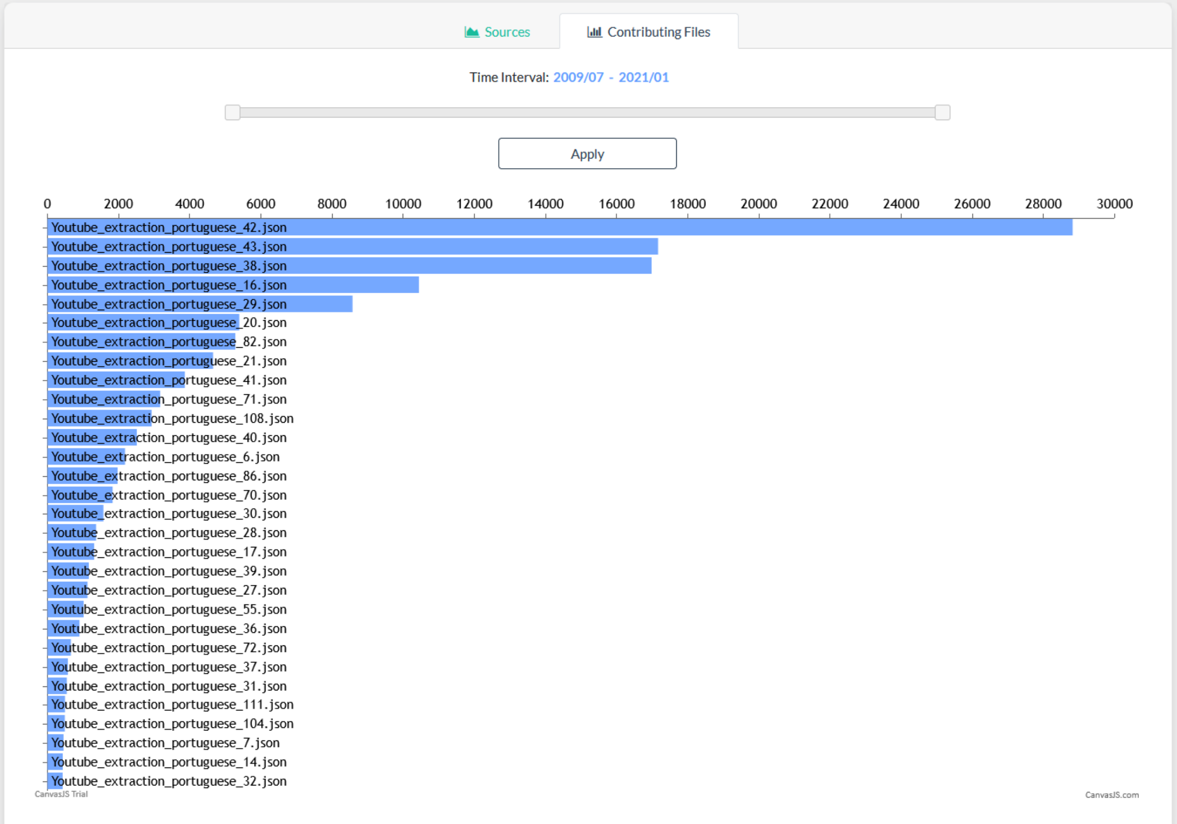

Sources
Contributing Files
Time Interval: 2009/07 - 2021/01
Apply
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
24000
26000
28000
30000
Youtube_extraction_portuguese_42.json
Youtube_extraction_portuguese_43.json
Youtube_extraction_portuguese_38.json
Youtube_extraction_portuguese_16.json
Youtube_extraction_portuguese_29.json
Youtube_extraction_portuguese_20.json
Youtube_extraction_portuguese_82.json
Youtube_extraction_portuguese_21.json
Youtube_extraction_portuguese_41.json
Youtube_extraction_portuguese_71.json
Youtube_extraction_portuguese_108.json
Youtube_extraction_portuguese_40.json
Youtube_extraction_portuguese_6.json
Youtube_extraction_portuguese_86.json
Youtube_extraction_portuguese_70.json
Youtube_extraction_portuguese_30.json
Youtube_extraction_portuguese_28.json
Youtube_extraction_portuguese_17.json
Youtube_extraction_portuguese_39.json
Youtube_extraction_portuguese_27.json
Youtube_extraction_portuguese_55.json
Youtube_extraction_portuguese_36.json
Youtube_extraction_portuguese_72.json
Youtube_extraction_portuguese_37.json
Youtube_extraction_portuguese_31.json
Youtube_extraction_portuguese_111.json
Youtube_extraction_portuguese_104.json
Youtube_extraction_portuguese_7.json
Youtube_extraction_portuguese_14.json
Youtube_extraction_portuguese_32.json
CanvasJS Trial
CanvasJS.com

Figure 7: Image displaying the number of comments that each file contributed, during a certain time interval

The second graph (Figure 7) is presenting the same information as the last one, but while the previously mentioned *Sources* graph is grouping the comments according to their source of extraction, the *Contributing Files* graph is instead looking at their file of origin. In other words, for any given set of comments, the resulting graph is showing the list of files from where the set of comments was sampled from, with their respective comment count, represented by the horizontal blue bars. This list is then ordered in such a way so that the file with the most comments can always be found at the top, while the rest will follow in a descending order, up to a maximum of thirty file names, with their respective comment counts.

This graph also allows for the specification of a time interval that will be used to filter the current set of comments according to their date of creation. In order for the user to inform the platform which is the desired time interval, all that is necessary is to simply adjust the two slidable squares along the horizontal frame, located above the graph itself, as it is being showed by the Figure 7. After the appropriate time interval is defined, the information contained within the graph shown below will be updated with the new filter once the *Apply* button is pressed.

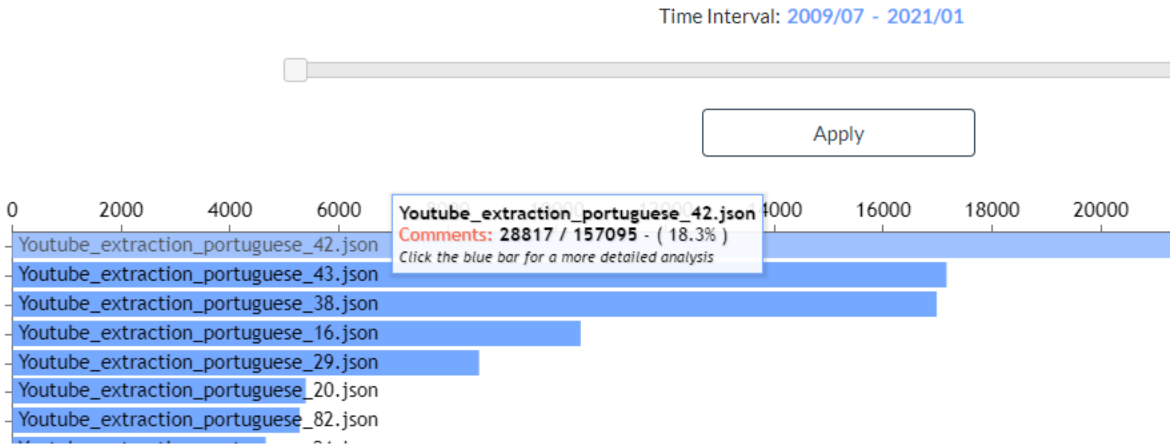

Time Interval: 2009/07 - 2021/01
Apply
0
2000
4000
6000
4000
16000
18000
20000
Youtube_extraction_portuguese_42.json
Youtube_extraction_portuguese_43.json
Youtube_extraction_portuguese_38.json
Youtube_extraction_portuguese_16.json
Youtube_extraction_portuguese_29.json
Youtube_extraction_portuguese_20.json
Youtube_extraction_portuguese_82.json
Youtube_extraction_portuguese_42.json
Comments: 28817 / 157095 - (18.3%)
Click the blue bar for a more detailed analysis

Figure 8: Image displaying the tooltip for contributing files graph

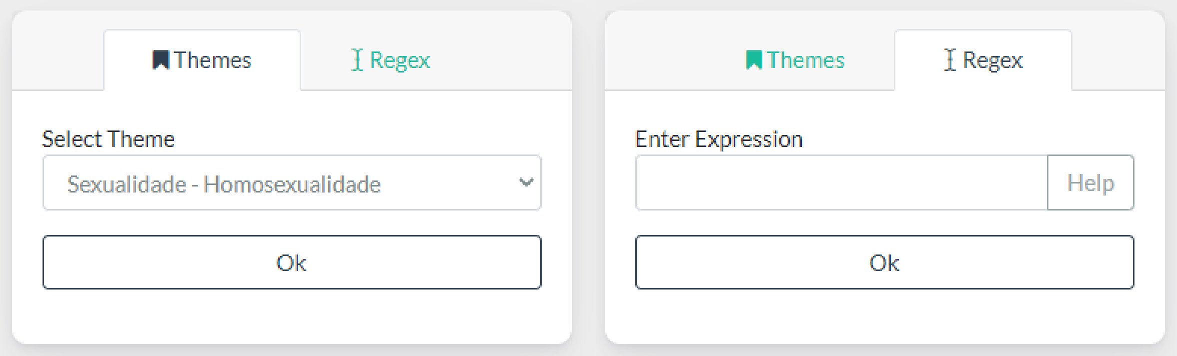
Furthermore, each one of the horizontal blue bars being displayed by this graph will present additional information regarding that file entry. When hovered over with the cursor, illustrated by the Figure 8 a small box will appear and display the exact comment count and the percentage of comments that are being provided by each one of the files listed, for the previously defined time interval.

Lastly, with the aim of providing a more intuitive and fluid browsing experience, the *Contributing Files* graph allows for each one of the blue bars to be clicked, which will in turn redirect the user on a new tab to the respective page of the *Search by File* mode [1], and display all of the information extracted from the file that the blue bar in question represents.

5.1.2 *Theme Search*

When choosing the *Search by Theme* option, the top of the page will display a small navigation menu with two independent search modes (Figure 9). The first one allows for one of the predefined themes to be chosen. The second one is designed for a more customizable search experience. It allows the introduction of regular expressions, which will be used by the engine to sift throw the complete set of comments, and find a subset that satisfies the search conditions.

[1] Further explained in Section 5.1.3



Select Theme: Sexualidade - Homosexualidade

Enter Expression:

Figure 9: Image displaying both options of the navigation menu from the *Search by Theme* page

Once either the theme or regular expression is submitted, the data extracted from the subset of comments is presented in different ways. At the center of the page, four graphs will be displayed: *Polarities*, *Sources*, *Theme vs NetLang* and *Contributing Files*.

The second and fourth graphs behave in a similar fashion to their same name counterparts, from the *Source Analysis* view mode (previously described in Section 5.1.1), the only difference being that instead of looking at all of the corpus' comments indiscriminately, they are only showing information regarding the current subset of comments.

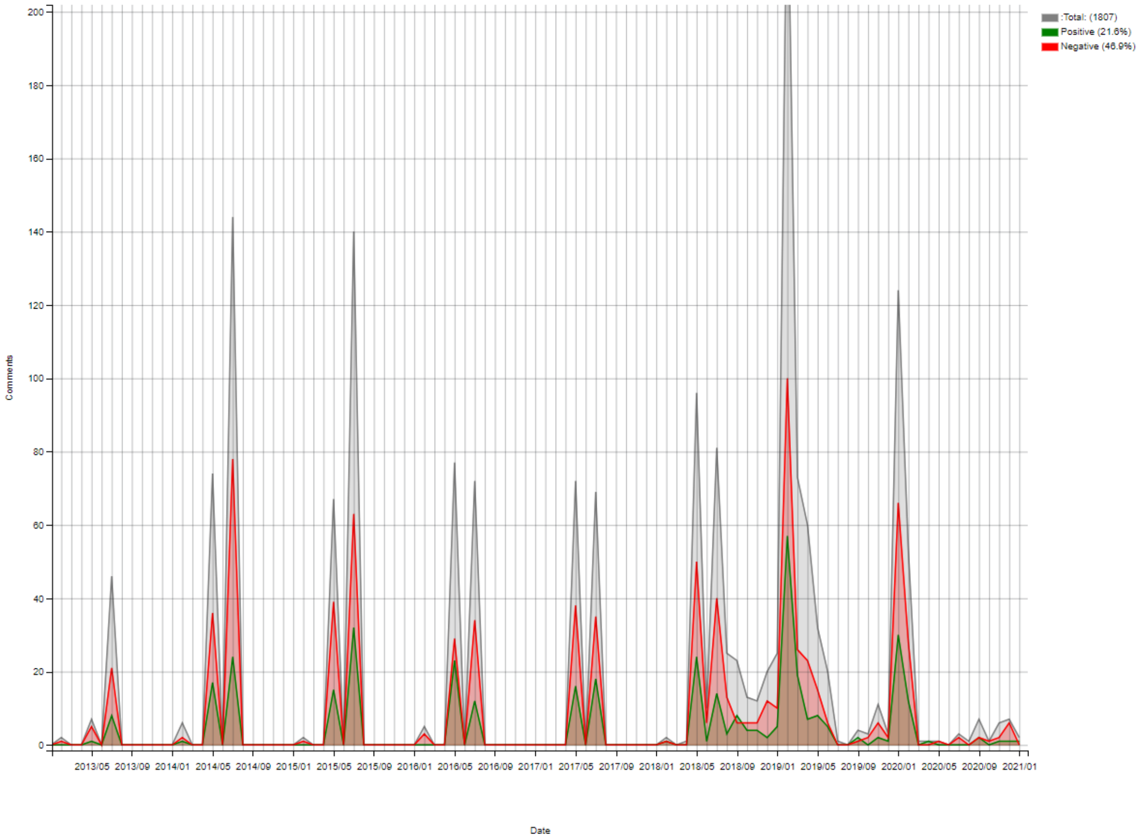


Figure 10: *Polarities* graph obtained by selecting the predefined theme of *Black Ethnicity*

The *Polarities* graph (Figure 10), separates the comments based on their polarity (green plot for positive and red one for negative comments), overlapping these two plots over the grayed out total subset.

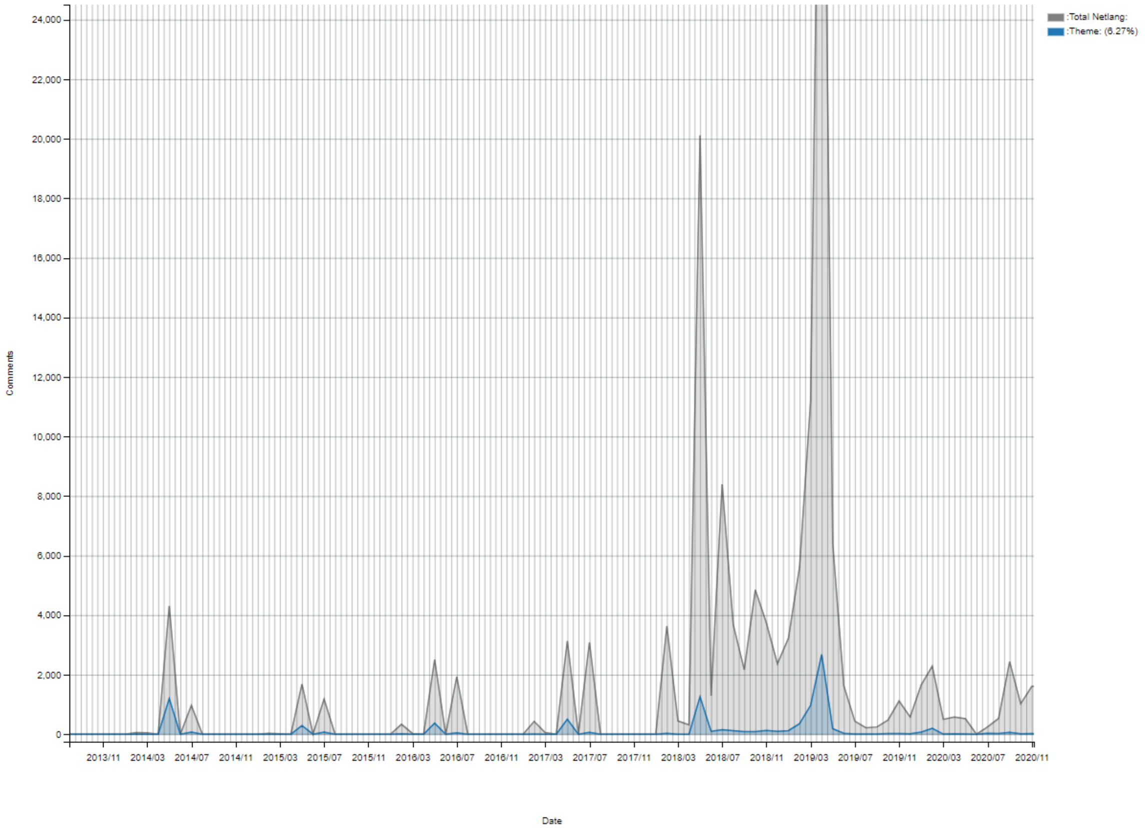
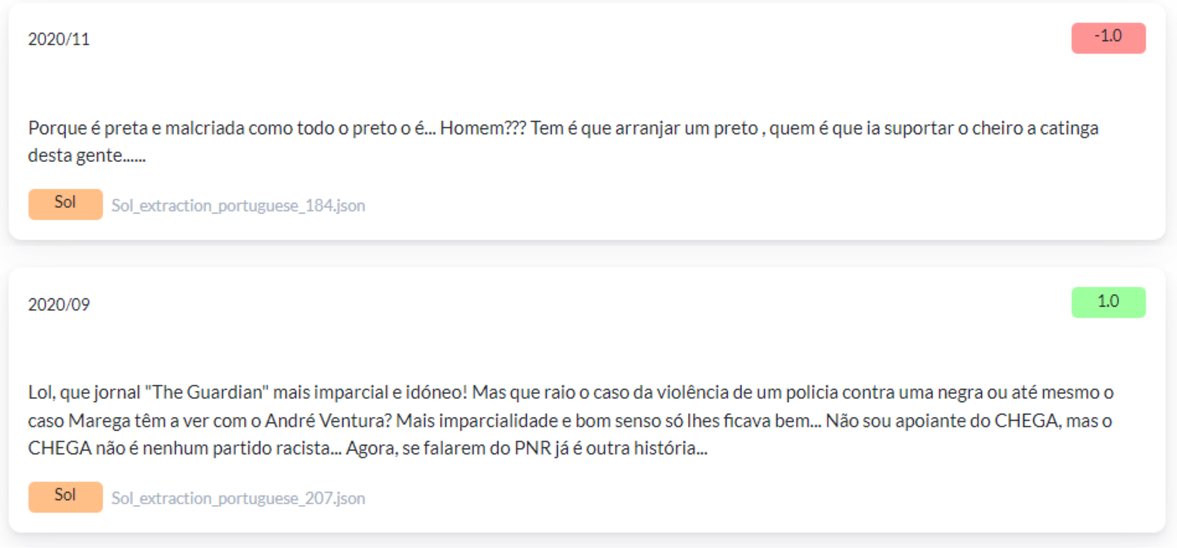


Figure 11: *Theme vs NetLang* graph from the *Black Ethnicity* theme page

Lastly, although it may have the least intuitive name of the four, what the *Theme vs NetLang* graph (Figure 11) shows is quite simple: it draws a visual comparison of total comments per month, between the current subset and the complete set of comments from the corpus. This comparison can be useful when trying to determine how widespread a certain topic is, within the corpus, possibly even helping to diagnose cases where the corpus is skewed towards a specific topic or theme.



2020/11

-1.0

Porque é preta e malcriada como todo o preto o é... Homem??? Tem é que arranjar um preto , quem é que ia suportar o cheiro a catinga desta gente......

Sol Sol_extraction_portuguese_184.json

2020/09

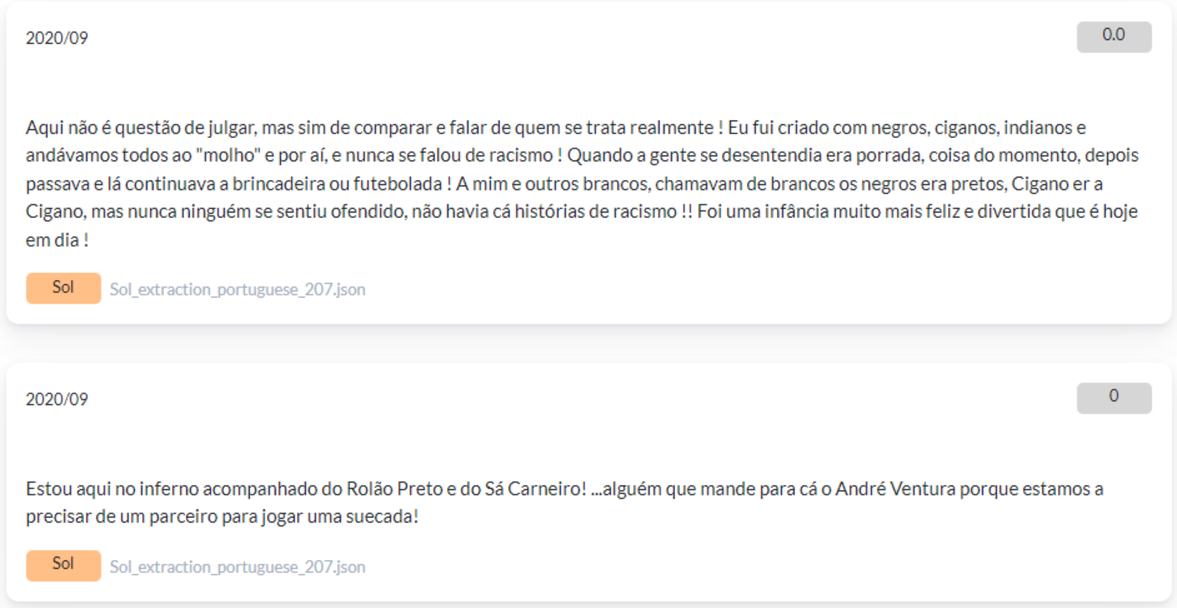
1.0

Lol, que jornal "The Guardian" mais imparcial e idóneo! Mas que raio o caso da violência de um policia contra uma negra ou até mesmo o caso Marega têm a ver com o André Ventura? Mais imparcialidade e bom senso só lhes ficava bem... Não sou apoiante do CHEGA, mas o CHEGA não é nenhum partido racista... Agora, se falarem do PNR já é outra história...

Sol Sol_extraction_portuguese_207.json

Figure 12: Example showing a positive and negative comment for the current theme

Underneath the different graphs, the list of comments being used by the current subset is displayed. Each entry shows information regarding that comment's creation date, the source of extraction, its polarity value and the actual written content of the comment. The polarity value is also color code: positive comments have green values and negative ones have red, as it can be seen in Figure 12.



2020/09

0.0

Aqui não é questão de julgar, mas sim de comparar e falar de quem se trata realmente ! Eu fui criado com negros, ciganos, indianos e andávamos todos ao "molho" e por aí, e nunca se falou de racismo ! Quando a gente se desentendia era porrada, coisa do momento, depois passava e lá continuava a brincadeira ou futebolada ! A mim e outros brancos, chamavam de brancos os negros era pretos, Cigano er a Cigano, mas nunca ninguém se sentiu ofendido, não havia cá histórias de racismo !! Foi uma infância muito mais feliz e divertida que é hoje em dia !

Sol Sol_extraction_portuguese_207.json

2020/09

0

Estou aqui no inferno acompanhado do Rolão Preto e do Sá Carneiro! ...alguém que mande para cá o André Ventura porque estamos a precisar de um parceiro para jogar uma suecada!

Sol Sol_extraction_portuguese_207.json

Figure 13: Example showing both cases of neutral comments for the current theme

For those cases when the polarity dictionary was unable to identify any kind of term that would indicate a polarity, the comment will be considered as neutral and its polarity value will be set to **0** (zero) and colored in grey, illustrated by Figure 13. Lastly, when the polarity dictionary identifies an equal number of positive and negative terms in a comment, its polarity value is set to **0.0** (zero point zero) and also colored in grey.

5.1.3 *File Search*

When the search mode *Search by File* is selected, the user will be taken to a page containing a small menu (Figure 14) that allows for one of the files that are being used by the platform to be selected.

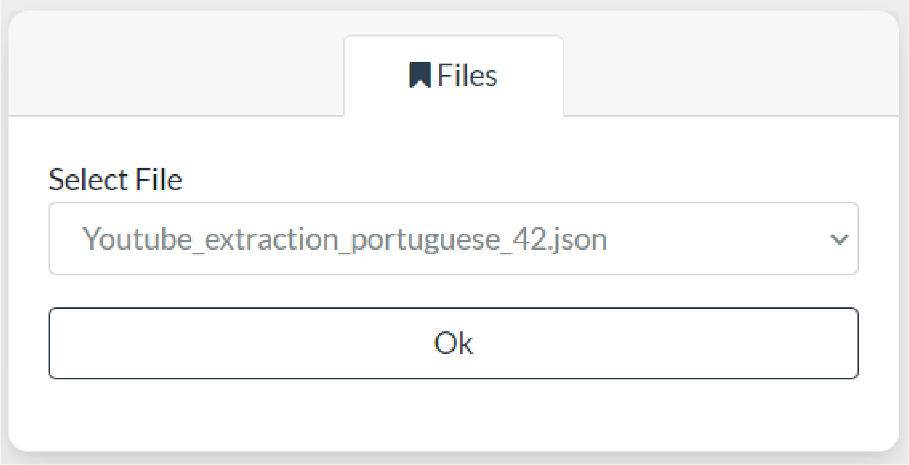
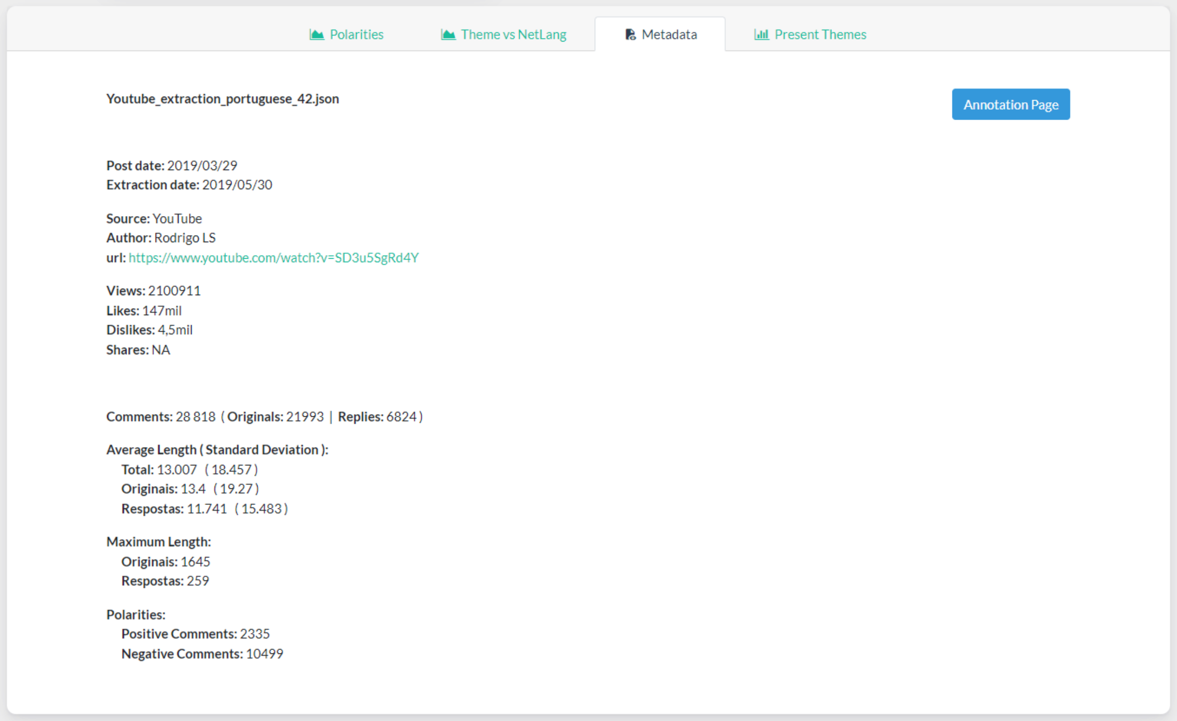


Figure 14: Image displaying the navigation menu from the *Search by File* page

Once a file has been selected, another page will be displayed, where four tabs are being exhibited at the top. These are the *Polarities* graph, the *Theme vs NetLang* graph, the *Metadata* tab and the *Present Themes* graph.

The first two of these graphs work in a similar fashion to their same name counterparts present in the previous search mode. The only difference being that, instead of looking at a sample of comments selected from the wider corpus, all of the statistics displayed within the *File Search* are being produced as a result of the analysis of the entirety of the comments that are contained by that specific file.



Youtube_extraction_portuguese_42.json

Post date: 2019/03/29
Extraction date: 2019/05/30

Source: YouTube
Author: Rodrigo LS
url: https://www.youtube.com/watch?v=SD3u5SgRd4Y

Views: 2100911
Likes: 147mil
Dislikes: 4,5mil
Shares: NA

Comments: 28 818 (**Originals:** 21993 | **Replies:** 6824)

Average Length (Standard Deviation):
Total: 13.007 (18.457)
Originais: 13.4 (19.27)
Respostas: 11.741 (15.483)

Maximum Length:
Originais: 1645
Respostas: 259

Polarities:
Positive Comments: 2335
Negative Comments: 10499

Figure 15: Image displaying the contents of the *Metadata* tab from the *Search by File* page

The third tab that can be found in the *File Search* mode, contains some useful metrics regarding the contents of the file in question, that can be helpful with the analysis process and shape the way one may arrive at a certain conclusion.

As it can be seen in Figure 15, the *Metadata* tab also displays some extra useful information, namely the link to the original post from where the file was originated.

It is also worth mentioning that a button can be found in the top right corner of this tab, that when pressed it will take the user to the corresponding file page on the *NetLangEd* platform, which is a web editor developed by Rui Rodrigues for the exact purpose of supporting online comment analysis and annotation (Rodrigues, 2022; Rodrigues et al., 2021).

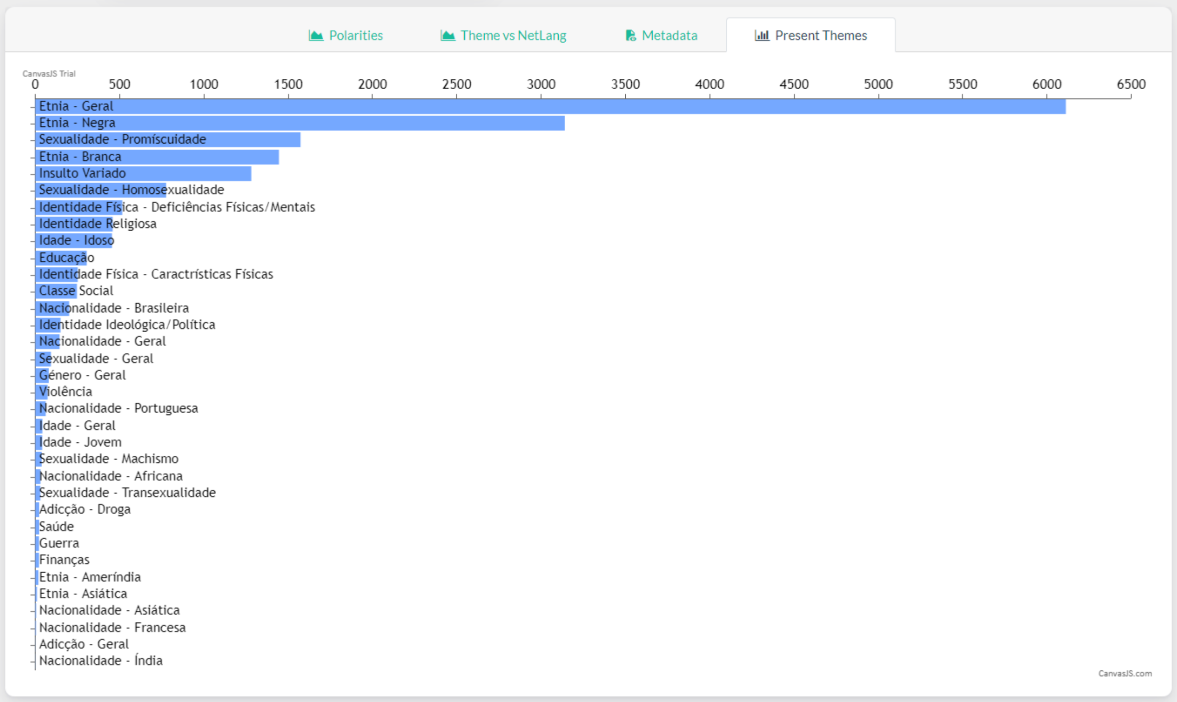
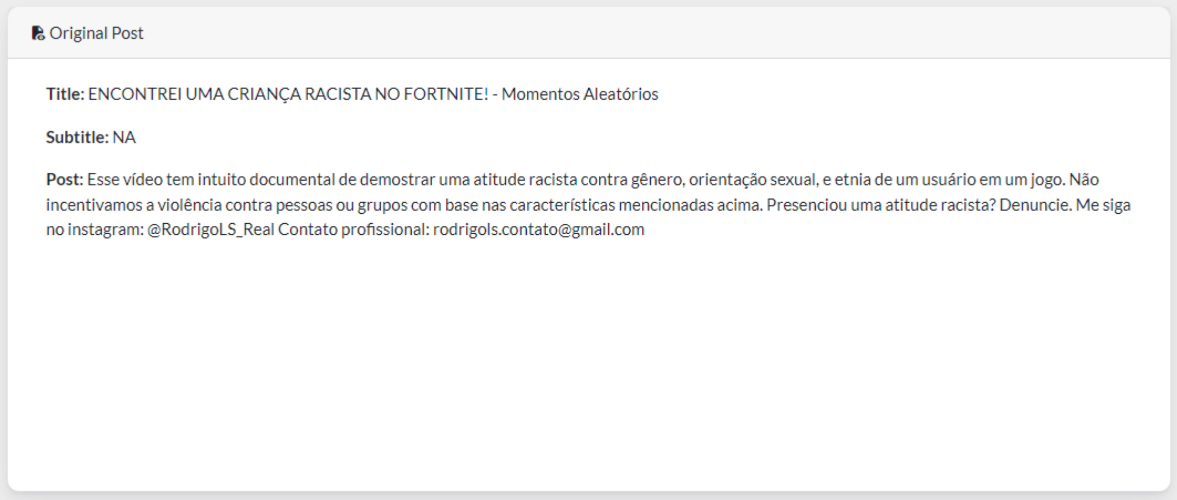


Figure 16: *Present Themes* graph from the *Youtube_extraction_portuguese_42.json* file page

The fourth and final tab of this search mode contains the *Present Themes* graph, illustrated by the Figure 16. This graph offers a visual representation of how many comments of that specific file were identified by the program as being a part of a certain theme.

Each one of the identified themes, represented by the horizontal blue bars, when clicked will open a new page and redirect the user to the respective *Theme Search* page and present the same information as it was previously previously described in Subsection 5.1.2.

This interconnectivity of the multiple search modes was intentionally designed in this manner, with the aim of providing the user with a more fluid and organic corpus and comment analysis experience.


Original Post

Title: ENCONTREI UMA CRIANÇA RACISTA NO FORTNITE! - Momentos Aleatórios

Subtitle: NA

Post: Esse vídeo tem intuito documental de demostrar uma atitude racista contra gênero, orientação sexual, e etnia de um usuário em um jogo. Não incentivamos a violência contra pessoas ou grupos com base nas características mencionadas acima. Presenciou uma atitude racista? Denuncie. Me siga no instagram: @RodrigoLS_Real Contato profissional: rodrigols.contato@gmail.com

Figure 17: Image displaying the written contents of the original post from the *Youtube_extraction_portuguese_42.json* file

Another novelty that can only be found in the *File Search* mode is the original post's information. When using this search mode, this section will appear located immediately underneath the four previously mentioned tabs, and above the complete list of comments contained within the file in question (this list will display information in a similar way as the one described in Subsection 5.1.2).

As it can be observed in Figure 17, the *Original Post* section will present more details regarding the title of the post, the subtitle, when one is present, and the post's written content. For the files extracted from *YouTube*, a platform that offers a medium with a more visual nature, the post's written content would be equivalent to the video's description.

5.2 INSIGHTS

The implementation process of this tool was one where there was a lot of trial and error, both in terms of how the information provided by the *NetLang Project* was processed, as well as the numerous experiments on how to best view the data and present the results to the final user. As a byproduct of this implementation process, since the data could be observed so closely and from so many unique perspectives, a few insights were allowed to be revealed, regarding the source material along with the tool itself.

5.2.1 *Tendency*

Firstly, by glancing through the total number of comments that each source (*YouTube*, *Publico* and *Sol*) is contributing towards this corpus, with the assistance of the tool's *Source Analysis* function [2], one is immediately able to understand that the three sources are not balanced. In fact, as it can be easily discerned from the previous Figure 6, as of the moment of writing of this document, the set of comments that constitute the *NetLang Corpus* is heavily biased towards the sourced *YouTube* platform, which is responsible for about 87% of individual comments.

Such a trend was somewhat expected, since the source in question is prone to have a higher reach and users tend to be actively encouraged by the content creators to leave their opinions about their video. Knowing this may help to justify why the *YouTube* comments are more numerous while being of lesser quality, for the most part, both when it comes to their content (there are a lot of insults thrown around) as well as their written structure.

Having said that, and although the previously offered explanation is still valid, the problem is more complex than that. By taking a closer look at Figure 7, it becomes apparent that 40% of all the *NetLang Corpus* comments are contained within just three files:

- Youtube_extraction_portuguese_42.json with 18.3%
- Youtube_extraction_portuguese_43.json with 10.9%
- Youtube_extraction_portuguese_38.json with 10.8%

This means that the corpus is not only biased towards *YouTube* as a source, but also that a small number of files from that source are responsible for a disproportionately large number of comment contributions to the *NetLang Corpus*. This poses a problem when it is time to draw conclusions or make predictions by using this data, because a minority of sources and files will have a deceptively large representation for certain themes and within their given time period. Now that this information has come to light, an analyst or researcher, that is intending on using this data in order to extract their corollaries, should keep this in mind in an effort to avoid making a hasty assessment about what they are trying to find within this corpus.

5.2.2 *Timestamps*

As it was previously explained in more detailed in Subsection 4.1.2, for a large portion of the comments, the exact creation date was unable to be successfully extracted. Since these cases were too numerous and simply discarding the would be a waste, a decision was made to preserve these cases and an attempt was made to predict their actual month of creation.

Since the current data contains comments with approximated dates, some of the temporal graphs, that are being generated by the platform, will not be capable of displaying accurate information regarding the time period of certain themes or trends. As a consequence of the

[2] Refer to Subsection 5.1.1

nature of the method that is being employed for the comment date prediction process, some of the temporal graphs may appear to arrange their data in such a way that a cyclical trend may be inferred. Until these dates are corrected, or the comments themselves are either removed or ignored, researchers that, at the time of writing, intend on using this platform with the current data, are advised to remain sceptical whenever they find themselves looking at an apparently cyclical graph.

A simple measure that users can take, in order to prevent themselves from quickly jumping to the conclusion that a certain theme graph has a cyclical nature, is to make use of the *Theme vs Netlang* tab and check if the spikes existing in the current theme graphs are aligned with the spikes of the wider unfiltered *NetLang* corpus.

5.2.3 *Graphic Interface*

The development process of this project as a whole took a very serendipitous approach.

Currently, the python library that is being used to generate many of the graphs that can be seen in the online platform is the *mpld3* library[3]. Its defining characteristic is that it can export the graphs that are being produced by the *matplotlib* library, directly into *html* and *JavaScript* code. Even before the idea of a web application was even formulated, there were already several graphs that were being generated with the help of the *matplotlib* library. When it was time to adapt the code in order to allow for these graphs to be displayed online, *mpld3*'s portability made the prospect of its use in the web platform very enticing, since it would mean that a large part of the already existing code would not need to be discarded and need to be implemented from scratch.

Initially, this decision proved to be a time saver, and looking back it can still be considered a good choice, since the main selling point remains true: the fact that the dataframes and already existing code used by the *matplotlib* library could be reused by the *mpld3* functions speed up the web interface's development process.

The main negative aspects from having this library be a part of this project are the fact the *mpld3* library is still a work in progress and, as one would expect, it still requires a considerable amount of work until it can be called a finished product. Currently, it contains from a few small visual bugs and inconsistencies, to *matplotlib* features that do not translate very well, and even some cases where entire features have no equivalent in the *mpld3* output, because the authors haven't had the time to fully implement them yet. In retrospect, what this library allows to be done is somewhat limited, therefore if another that offered more flexibility was chosen in its place, perhaps the resulting web application would have been one of higher quality.

[3] https://mpld3.github.io/

As a consequence of the aforementioned issues, a decision was made to overhaul the online platform with a different library, at a later date when this project is revisited. This will allow for the improvement of the current available features, in conjunction with the experimentation of previously unexplored ones. With this approach the final product will be enhanced, although it will mean that a large portion of the code will have to be discarded, which was what the initial decision of using the *mpld3* library was attempting to avoid.

5.3 SUMMARY

The main focus of this Chapter lies on describing the implemented web application in its entirety. A detailed explanation regarding the inner workings of the multiple search modes is being provided, as well as how to make optimal use of their respective features, in order to have the best possible user experience when exploring the provided data.

This Chapter concludes by presenting a few insights about the data and the tool itself, that were revealed during the implementation process of the web platform, and their possible implications are then discussed.

6

SOBA, VALIDATION

The moment the web application was finished, the decision was made that it would be necessary to validate the interface through an experiment with multiple users. This chapter focuses on describing this experimentation and analysing its results.

6.1 FINAL USER EXPERIMENT

During the *NetLang Project*'s Autumn Workshop on "Online Hate Speech: A Corpus Linguistics Approach", that took place on 11-12 of November 2021 at the University of Minho[1], multiple corpus exploration tools were presented, *SoBA* being one of them.

Starting with a brief initial presentation, followed by a more thorough and practical demonstration on how to best make use of the developed platform, the attendants of the aforementioned event were given a closer look on how this tool could be of assistance to their process of corpus analysis. During the afternoon, with the aim of being able to measure how successful the implemented approach was, the participants were given a testing guide (described in appendix A), that was previously designed, as an assignment. Once they had completed the exercise, a link to a questionnaire (described in appendix B) was provided with the goal of gathering data on the tool's performance, as well as other metrics and suggestions on how further improvement could be achieved.

Afterwords, the same assignment and questionnaire were given to both groups of students and other people that were not related to this area of knowledge, as a way of gathering more data and paint a broader picture of how this tool can be perceived by the general user.

6.2 ONLINE-SOBA, USABILITY QUESTIONNAIRE RESULTS

The idea behind using a questionnaire was so that it would be possible to gather feedback regarding the ease of use of the developed platform, as well as to get a better grasp on what crucial design changes needed to be made.

[1] A photo gallery of the event can be found at their website

This questionnaire can be divided in two parts. The first is used to gather information about what kind of people were following the testing guide and what their academic backgrounds were. The second part contains more specific questions regarding the user's experience when using the site, while also providing some information that may help identify possible shortcomings within itself.

6.2.1 *User Identification*

The number of people that decided to complete the last step of the testing guide and follow through with the usability questionnaire totaled at thirteen.

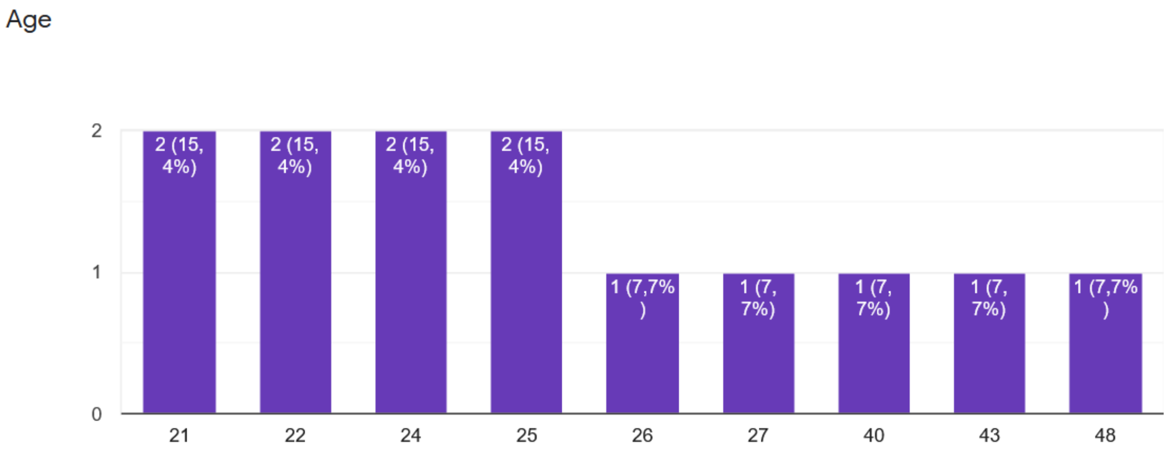


Figure 18: Bar chart displaying participant age distribution

After a brief overview of the age distribution chart (Figure 18), one can easily see that this set of ages can be organized into two age groups: *twenties* and *forties*. The entries from second group were authored by the few researchers from the *NetLang Project* that generously decided to contribute to the process of testing and improving of this tool. In contrast, the ones from the first group mostly came from other students enrolled at the University of Minho.

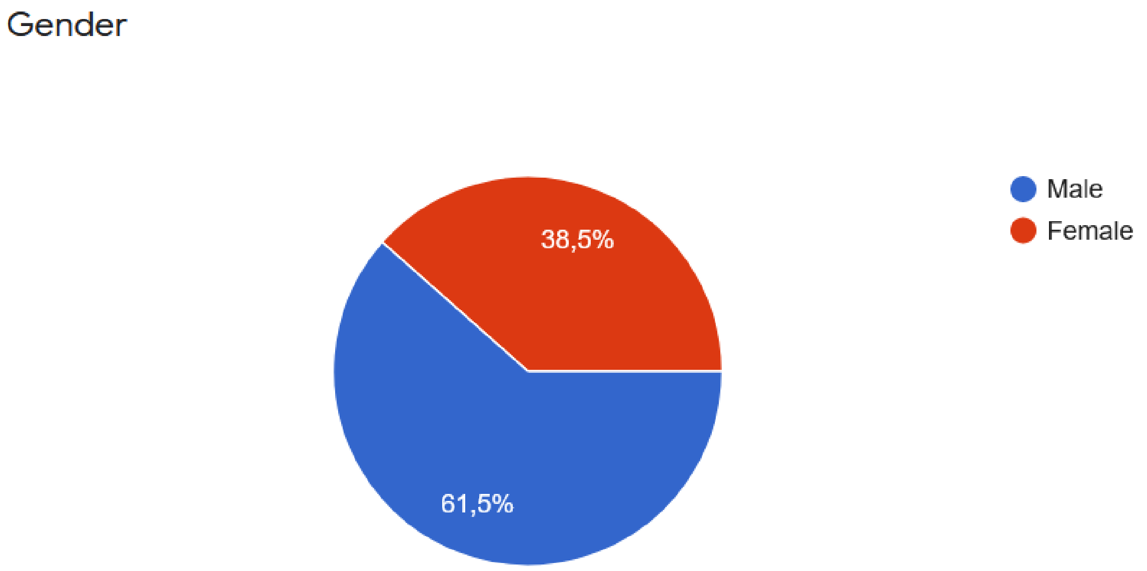


Figure 19: Pie chart displaying participant gender distribution

The participants were then asked what gender they identified themselves as. Around 60% identified as male, while the remaining ones identified themselves as female (Figure 19).

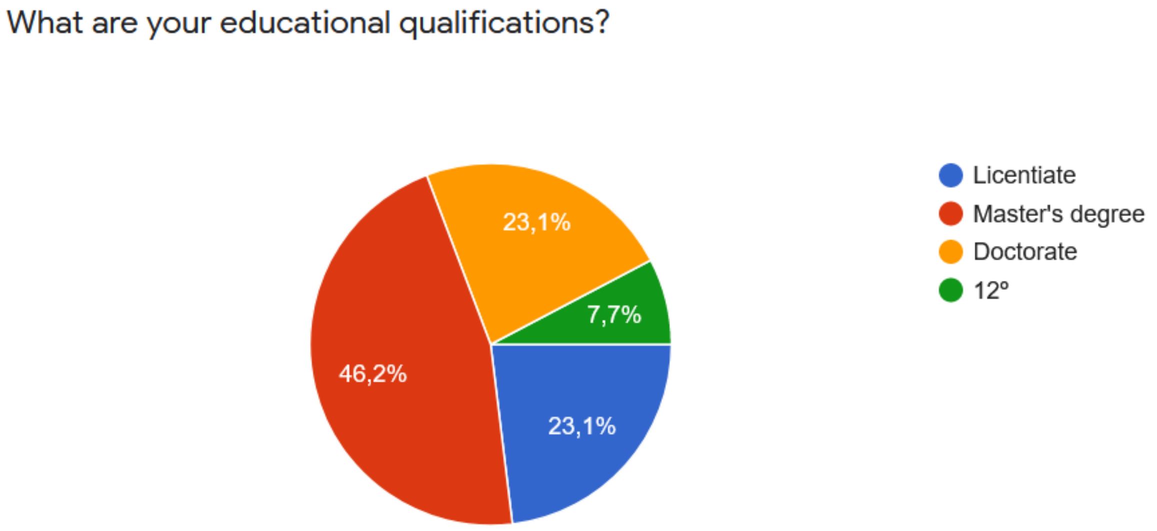


Figure 20: Pie chart displaying participant educational qualification

When looking at their qualifications, the majority of the participants had obtained at least a Master’s degree, with the three researchers possessing a Doctorate in their respective fields (Figure 20).

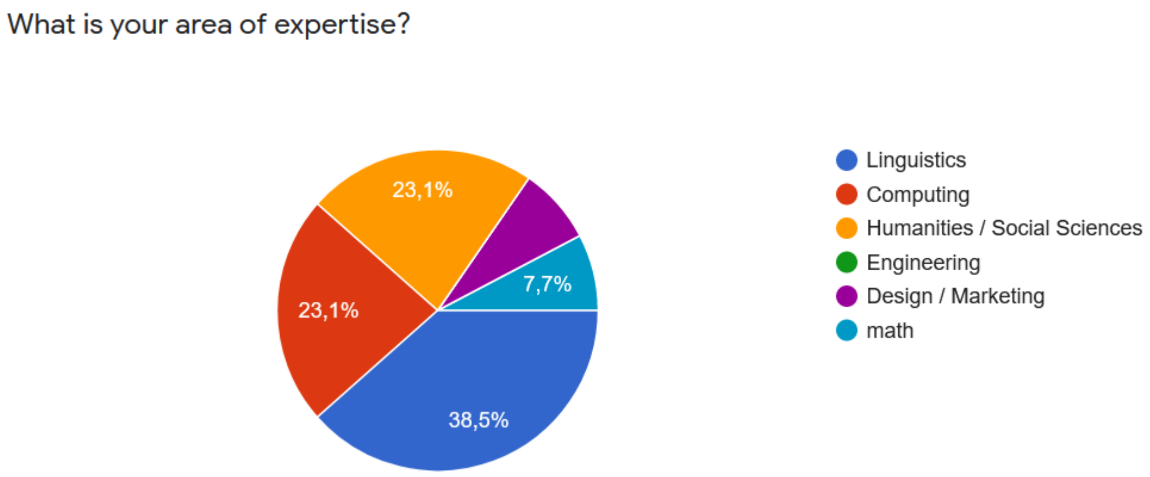


Figure 21: Pie chart displaying participant area of expertise

Regarding their area of expertise, a considerable portion of the thirteen participants come from a background in Linguistics, closely followed by Social Sciences and Computing (Figure 21). Considering the target audience for this tool is mostly researchers from the first two of these fields, means that the results of this modest study can be indicative of this tool's actual performance when used by professionals of these respective areas of knowledge.

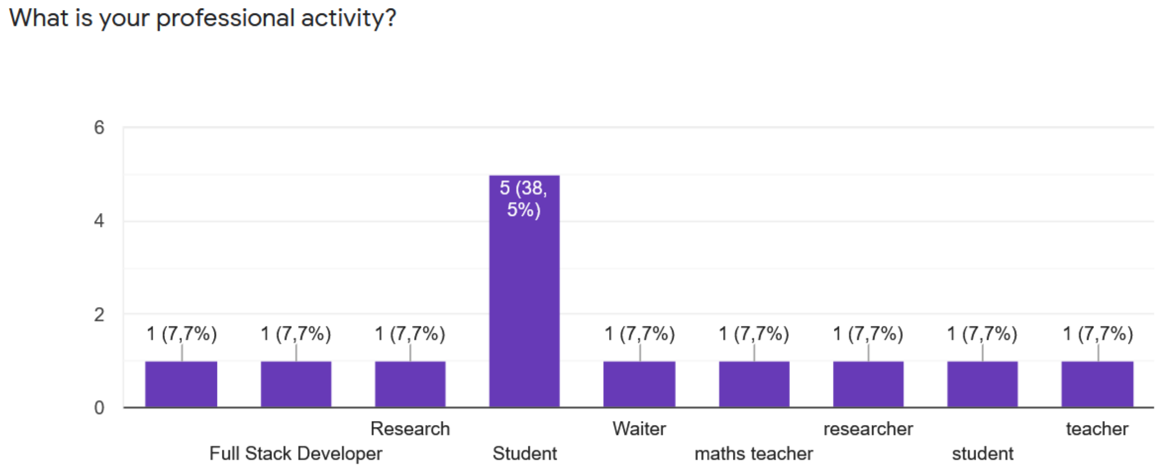


Figure 22: Bar chart displaying participant professional activity distribution

When asked about their professional activity, around half of the participants admitted to be students (Figure 22). This may indicate that our sample group of test subjects is somewhat inexperienced on the matter at hands.

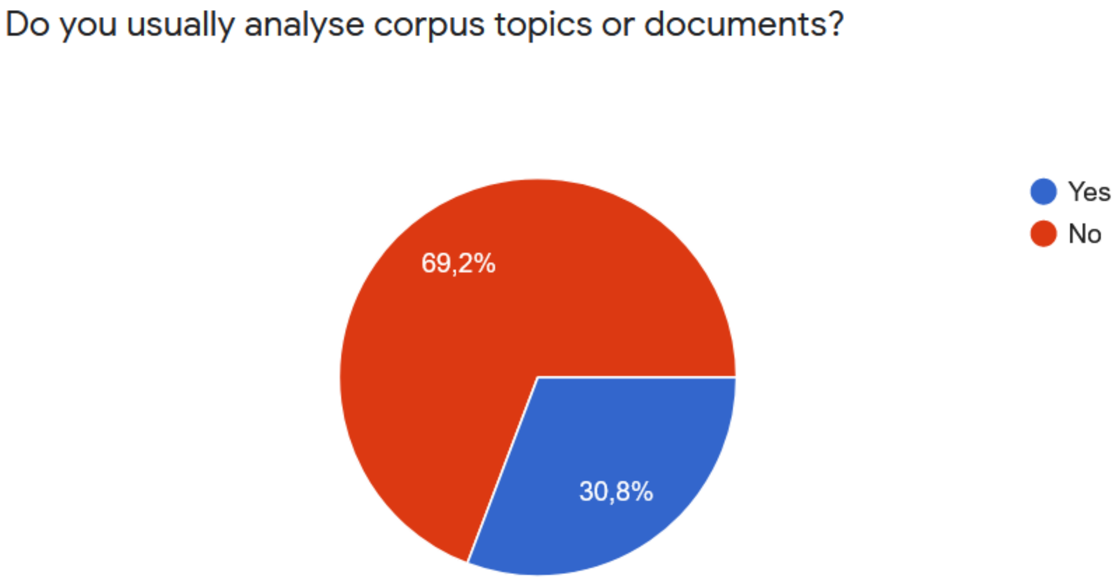


Figure 23: Pie chart displaying participant's familiarity with corpus analysis

When asked if they had any previous experience with corpus analysis, only about a third of the participants answered affirmatively (Figure 23). This is to be expected since more than half of the provided answers originated from students or people from different fields of knowledge, where the act of corpus analysis does not usually come up.

6.2.2 *User Experience*

As previously mention, this section of the questionnaire will be focusing more on how user-friendly the platform is and what could be some of the major improvement points.

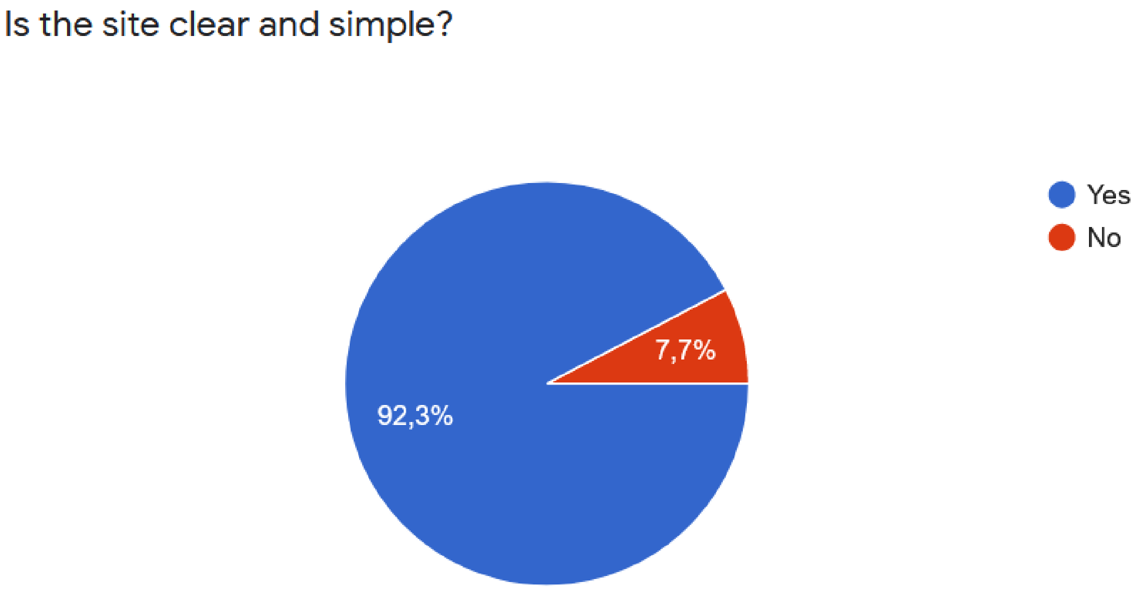


Figure 24: Pie chart displaying participant's opinion on whether the site was clear

User feedback regarding the site's current level of complexity and ease of use was very promising. It appears that only one participant thought that the platform was too complex (Figure 24).

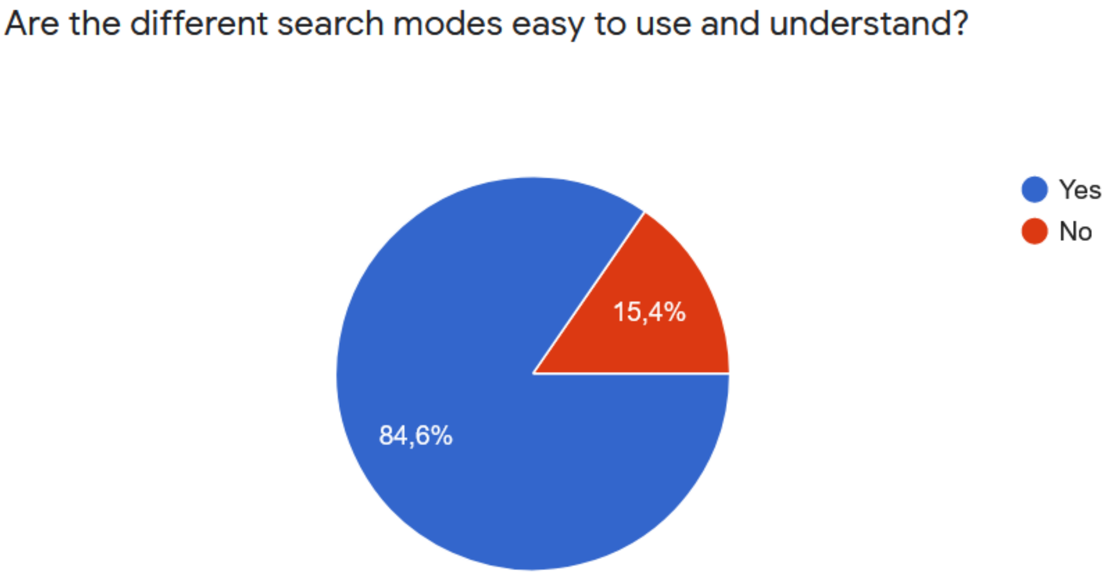


Figure 25: Pie chart displaying participant's opinion on whether the different search modes were easy to use and understand

As a followup to the previous question, the participants were asked on how they felt about the level of complexity of the different search modes. This was again answered in a very positive manner (Figure 25), although with slightly worse results than the previous, less specific question. This could be indicative that the different search modes need to be better explained to the user.

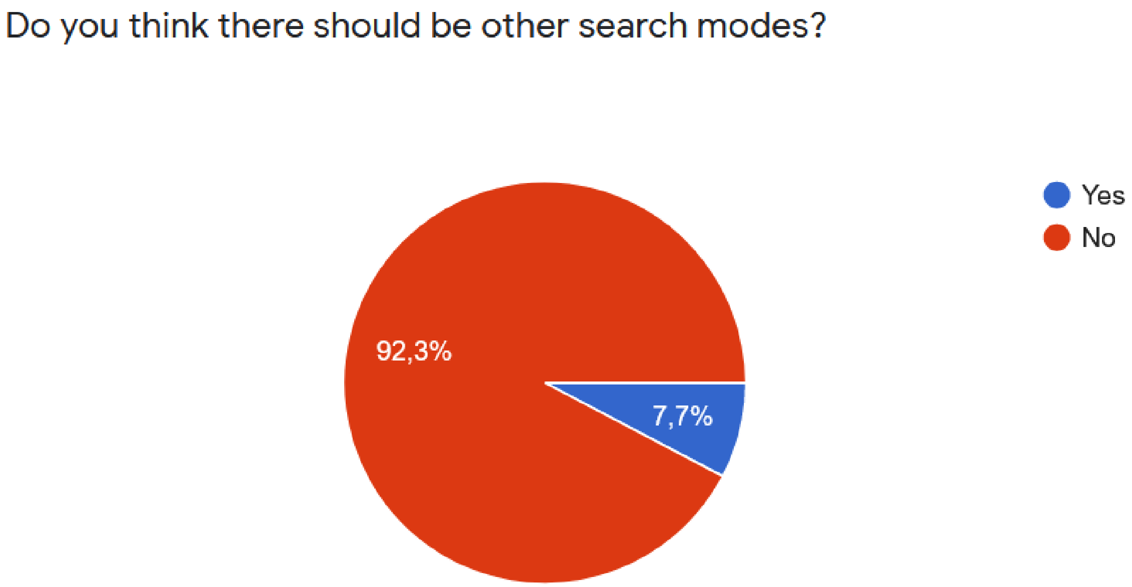
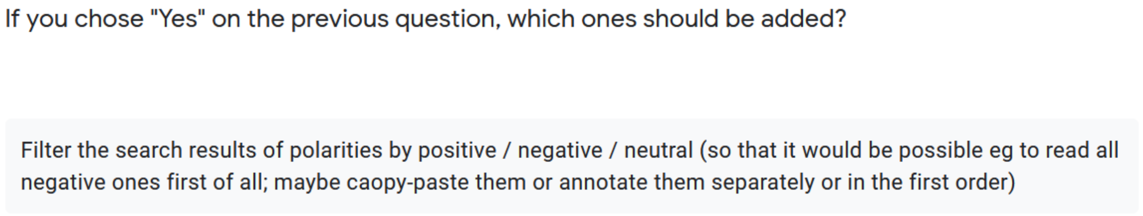


Figure 26: Pie chart displaying participant's opinion on whether new search modes should be added



If you chose "Yes" on the previous question, which ones should be added?

Filter the search results of polarities by positive / negative / neutral (so that it would be possible eg to read all negative ones first of all; maybe caopy-paste them or annotate them separately or in the first order)

Figure 27: Participant's suggestion for search mode improvement

The one participant that thought the site's usability could be improved with the addition of new search modes (Figure 26), decided to further elaborate their stance and provide us with a description of their idea (Figure 27). In essence, they recommend that whenever a list of comments is displayed, the ability to filter the provided results according to the polarity values of the comments. This is a wonderful proposal of improvement and it will be one of the main features added once the development of the tool is revisited in the near future.

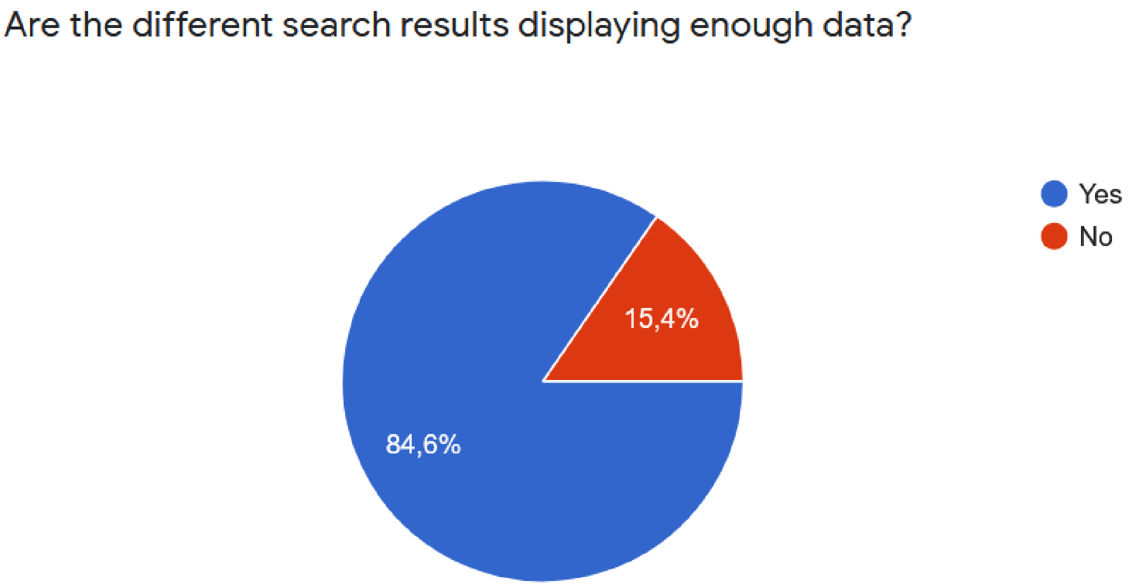
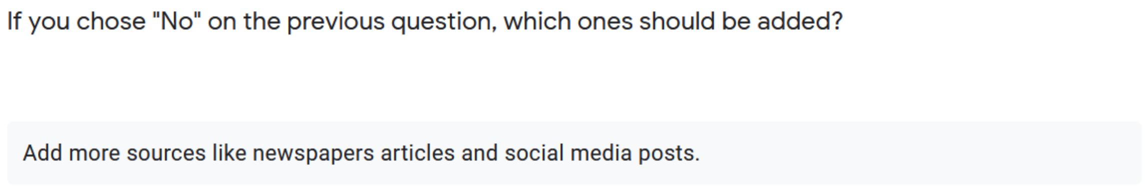


Figure 28: Pie chart displaying participant's opinion on whether the current search results are displaying enough data

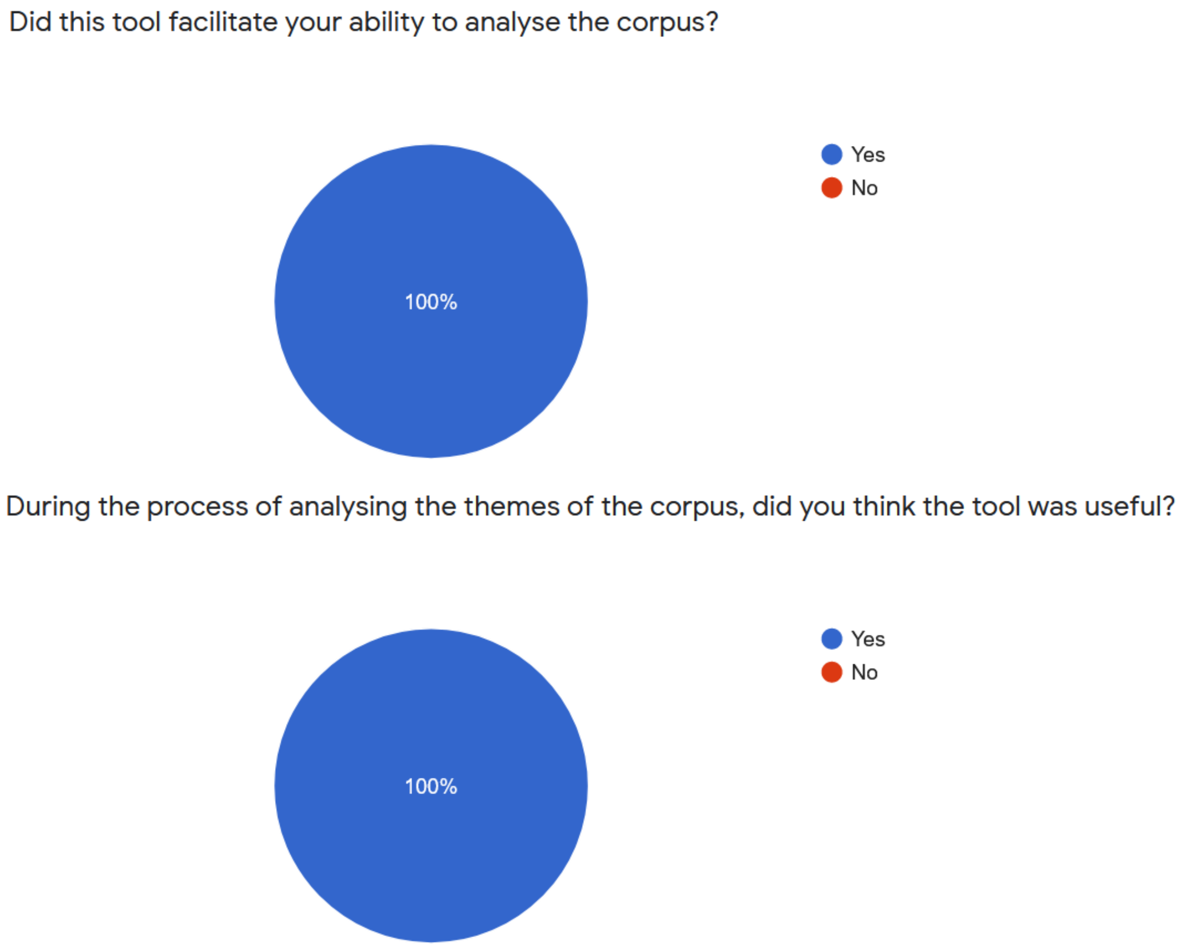


If you chose "No" on the previous question, which ones should be added?

Add more sources like newspapers articles and social media posts.

Figure 29: Participant's suggestion for improving the search results

Of the two participants that considered the current search results to be unsatisfying (Figure 28), only one of them decided to further elaborate their opinion by suggesting that other sources should be added to the platform (Figure 29). Although this is an excellent suggestion, since this project sourcing it's data from the *NetLang* Project, it means that this change would fall into the hands of the researchers responsible for it.



Did this tool facilitate your ability to analyse the corpus?

During the process of analysing the themes of the corpus, did you think the tool was useful?

Figure 30: Participant's general opinion on whether the tool was helpful during the analysis process

The participants were also questioned on whether the tool helped them analysing the corpus, or if when analysing the different themes, it proved to be useful. Their response to these questions could not have been more positive, since all of the users that decided to participate on this experiment then went on to answer affirmatively to both of them (Figure 30).

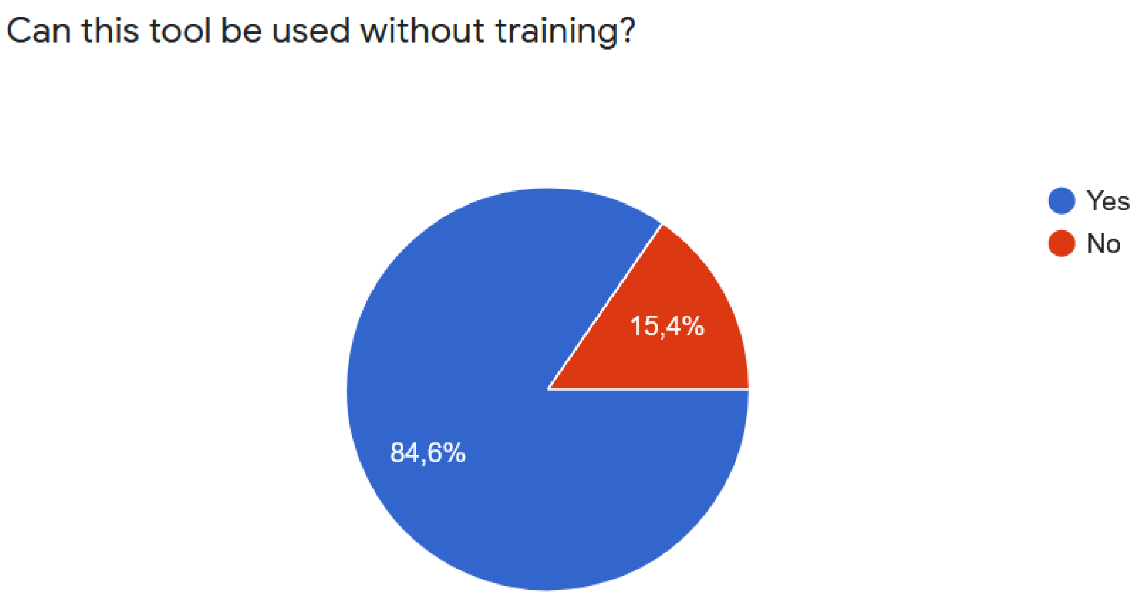
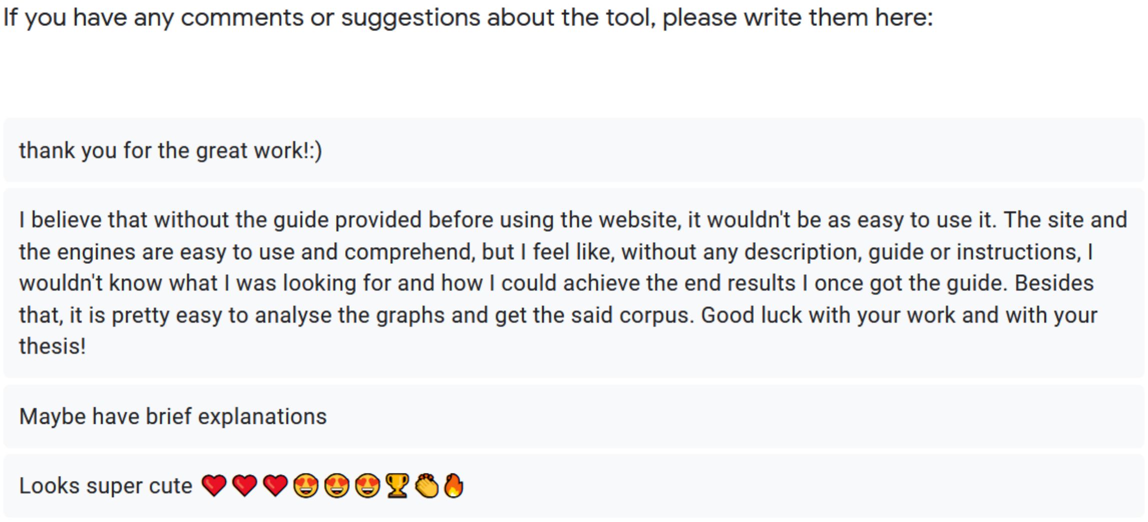


Figure 31: Participant's opinion regarding the feasibility of using this tool without any previous training

When asked, the overwhelming majority of the participants answered that this tool could be used without training (Figure 31). Considering the fact that only a small number of the test subjects had any previous experience with the process of analysing a corpus, this result is very promising, since it indicates that the platform is straightforward and intuitive enough to be used by mostly everyone from many different academic backgrounds.


If you have any comments or suggestions about the tool, please write them here:

thank you for the great work!:)

I believe that without the guide provided before using the website, it wouldn't be as easy to use it. The site and the engines are easy to use and comprehend, but I feel like, without any description, guide or instructions, I wouldn't know what I was looking for and how I could achieve the end results I once got the guide. Besides that, it is pretty easy to analyse the graphs and get the said corpus. Good luck with your work and with your thesis!

Maybe have brief explanations

Looks super cute ❤️❤️❤️😍😍😍🏆👏🔥

Figure 32: Set of comments and suggestions provided by the participants about the tool

Finally, on the last entry of the questionnaire, the users were invited to leave any suggestions that in their view could help with the further development and improvement of the platform.

There were four people that decided to share their ideas, as it can be seen in Figure 32. Of those four, two of them thought they should leave some words of encouragement for the work done, and those were very well received by the authors. The other two however, decided to leave some key points of improvement, which mainly revolved around making the site more clear for someone who would be using it for the first time. This suggestion has been thoroughly registered and will be one of the main focus when the time comes to do some restructuring work on the platform, where this and other changes will be implemented.

7

CONCLUSION

With the increase in the use of social networks and online communication, a vast amount of information is being produced every day by its user base. These platforms represent potential sources of behavioral information, not only about these individuals but about social groups as a whole.

In the context of this thesis, and through the application of techniques such as Natural Language Processing and Sentiment Analysis, it was possible to develop a tool capable of using this information in order to help researchers detect social trends and build a model of how they change over time.

A discussion on different approaches and attempts at solving the several obstacles that appeared during the development process has been provided, as well as the outcome of their application on the given test data, and what sort of compromises had to be made.

The main focus of this thesis was on the high level aspect of the development of a tool of this magnitude. Meaning that, more often than not, the individual components of the application were only thoroughly developed and fine tuned to the point were they were providing satisfactory results before moving to the next order of business. This approach was to guarantee that a functioning tool could be produced within the available time and with the use of the existing resources.

From an experimental point of view, the contributions of the work done in this thesis and the proposed solution are mainly two. Firstly, it provides a comparison between how different approaches, using Natural Language Processing techniques, fair against the challenge of implementing an application with this level of ambition to this kind of target data (comments extracted from online social media platforms), which is often characterized by its lack of lexical coherence and grammatical structure, which often is problematic for most classical approaches.

Secondly, it presents a possible answer to the question of: once all of the data is thoroughly sifted through, processed and automatically categorized, what would be the best way to display the application's findings to the final user. The currently existing web platform is this author's confident response to that problem, with the multiple navigational functionalities and several ways of filtering and analysing the given data.

The fact that during the final user experiment, described in more detail in chapter 6, the feedback provided by the researchers that are personally involved in the *NetLang Project* was extremely positive, further reinforces the notion that this attempt was a step in the right direction.

In the pursuit of truth, scientific research is a never ending-process of discovery. As such, the application developed in the context of this Master's Thesis, even though it is an extremely useful and functional tool, it cannot be considered a finished product and there is still a lot more work to be done.

7.1 FUTURE WORK

Many different areas of the application (some large, some small) can or are required to be further improved upon. Multiple adjustments, tests and experiments have been left for the future due to a lack of time. In order to allow the application to reach its full potential, it is necessary to restructure and optimise some of its components.

In the pre-processing pipeline, a different part-of-speech tagging tool is needed to replace *SpaCy*. The *FreeLing* language analysis tool seems to be a good candidate for experimentation, since it is being used by the *NetLang Query Engine* and has produced some very promising results Pereira (2022). Once this is achieved, it will be possible to improve both the spellchecker and the polarity identifier. Furthermore, the current theme identifier could do with some fine-tuning, as well as be expanded to encompass newer themes and a larger key-word count per theme.

Regarding the web platform, there are few changes and several features that would be beneficial if they were implemented in a future version of the application. For the reasons detailed in Section 5.2.3, the library that is currently being used to generate the temporal graphs needs to be replaced.

The rest of the proposed changes are focused on enriching the user experience in general, while simultaneously enhancing the tool's capability of thoroughly exploring the data:

- Whenever a user clicks on the file source of a comment, redirect them on a new tab to the corresponding *Search by File* page;

- For each comment, display the list of all the themes in which it has been identified;

- In the *Comments Used* section, allow the comments to be searched by a specific term or regular expression;

- In the *Comments Used* section, add the functionality to filter the comments by theme, source of extraction, file name, date of publication and date of extraction.

Finally, in order to validate the new version of the application, the final user experiment would have to be repeated in a similar fashion to the one described in Chapter 6, preferably with a higher number of test subjects.

A

ONLINE-SOBA TESTING GUIDE

A.1 CONTEXTUALIZATION:

This project, *Online-SoBA (Online Social Behavior Analysis)*, is a semi-automatic system capable of analyzing short texts corresponding to comments written in Portuguese in reaction to *'posts'* on social media platforms (namely social networks and online newspapers), so that it is possible to identify behaviors that define social opinions at a given time.

To face this challenge, the texts contained in the NetLang Corpus are analyzed, extracting information on the topic and the polarity of each one.

The calculation of the polarity of a comment is carried out through the number of words with positive or negative connotation that constitutes it. Subsequently, according to the sentence structures that characterize these behaviors in natural language, the theme of the comment is assigned.

By analyzing the resulting graphs of this process, we can quickly make a visual deduction about how social opinion on an issue has evolved over time.

A.2 FIRST PHASE (SOURCES):

- Go to Online Social Behavior Analysis;
- On the top of the page, select *Source Analysis*, to view the distribution graph of comments, grouped by the respective source of extraction;
- Draw your conclusions by manipulating the graph (it is possible to drag/zoom with the tools in the lower left corner, as well as disable each source individually by clicking on the graph's legend);
- Click on *Contributing Files* and set the *Time Interval* between 2020/01 and 2020/08;

- Click on the name of the first file for a more detailed view of it (this is similar to using the *Search by File* mode).

A.3 SECOND PHASE (THEMES):

1. Predefined Themes:
    - Select *Search by Theme* at the top of the page, to see information about the comments previously identified as containing a certain theme;
    - From the dropdown menu, select *Sexualidade – Homosexualidade*;
    - Explore the 3 graphs (*Polarities*, *Sources*, *Theme vs NetLang*) containing information on the topic under analysis and draw your conclusions (when in doubt, all graphs have tooltips with short descriptions of their content);
    - Click *Contributing Files* and set a time range so that you can see which files contributed the most comments in that range.
2. Custom Themes:
    - To change the search mode and look for a theme that does not exist yet, in the menu where you selected the theme you will click the search mode *Regex*;
    - Enter the regular expression: *andr.∗ventura*;
    - Notice that the *Regex* search mode presents results in a similar way to the previous one but has the advantage of allowing you to define your own themes through regular expressions;
    - Explore the different graphs and draw your own conclusions.

A.4 THIRD PHASE (FILES):

- Select at the top of the page *Search by File*, to see information about each file that is contained in the corpus;
- From the dropdown menu select *Sol_extraction_portuguese_145.json*;
- Explore the 4 modes of viewing information about this file (note that clicking on a theme in the *Present Themes* graph is similar to using the *Search by Theme* mode);

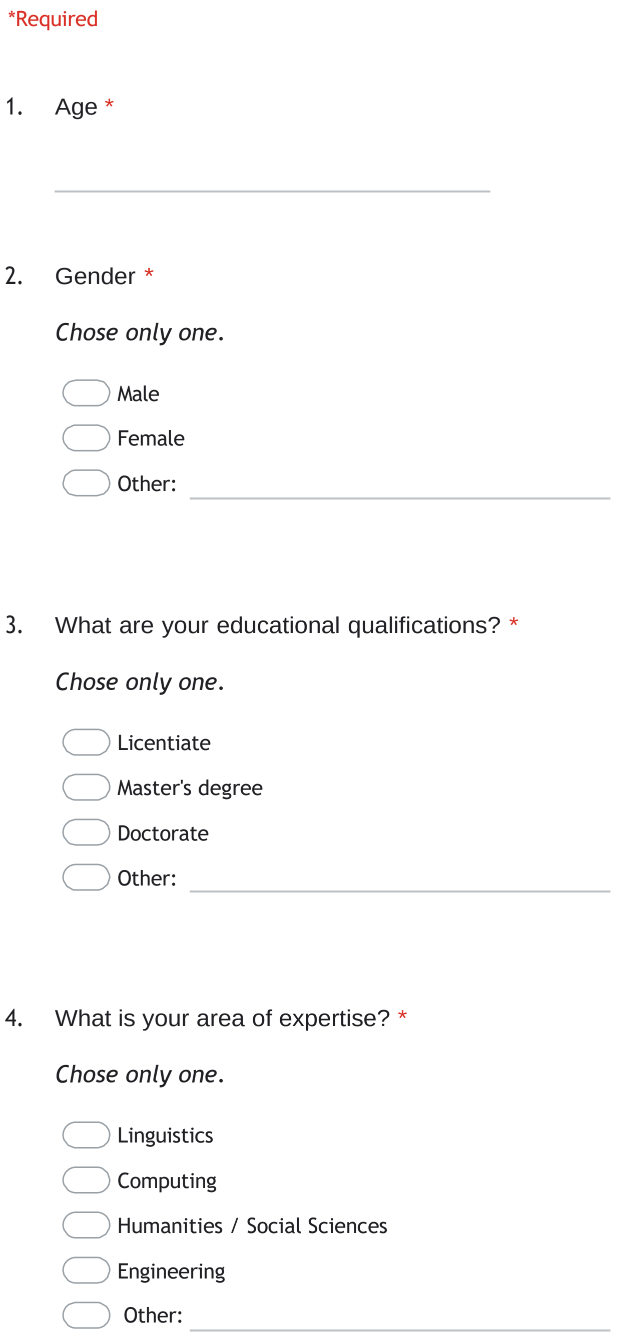
- After you have made a quick analysis of the original post's content and seen some of the comments, ask yourself whether the main themes being detected by the program make sense in the given context.

A.5 FOURTH PHASE:

- Answer the questionnaire available at `https://forms.gle/DKJdgVEYAWk337VL9`

B

ONLINE SOBA USABILITY QUESTIONNAIRE



*Required

1. Age *

____

2. Gender *

Chose only one.

○ Male

○ Female

○ Other: ________________________________

3. What are your educational qualifications? *

Chose only one.

○ Licentiate

○ Master's degree

○ Doctorate

○ Other: ________________________________

4. What is your area of expertise? *

Chose only one.

○ Linguistics

○ Computing

○ Humanities / Social Sciences

○ Engineering

○ Other: ________________________________

5. What is your professional activity? *

____

6. Do you usually analyze corpus topics or documents? *

Chose only one.

○ Yes
○ No

7. Is the site clear and simple? *

Chose only one.

○ Yes
○ No

8. Are the different search modes easy to use and understand? *

Chose only one.

○ Yes
○ No

9. Do you think there should be other search modes? *

Chose only one.

○ Yes
○ No

10. If you chose "Yes" on the previous question, which ones should be added?

11. Are the different search results displaying enough data? *

Chose only one.

○ Yes

○ No

12. If you chose "No" on the previous question, which ones should be added?

13. Did this tool facilitate your ability to analyze the corpus? *

Chose only one.

○ Yes

○ No

14. Can this tool be used without training? *

Chose only one.

○ Yes

○ No

15. During the process of analyzing the themes of the corpus, did you think the tool was useful? *

Chose only one.

○ Yes

○ No

16. If you have any comments or suggestions about the tool, please write them here:

BIBLIOGRAPHY

Cambridge_Dictionary. Hate Speech. https://dictionary.cambridge.org/us/dictionary/english/hate-speech. Accessed: 2020-10-09.

Paula Carvalho and Mário J Silva. SentiLex-PT 02. https://b2share.eudat.eu/records/93ab120efdaa4662baec6adee8e7585f. Accessed: 2020-12-14.

M. Dermouche, J. Velcin, L. Khouas, and S. Loudcher. A joint model for topic-sentiment evolution over time. In *2014 IEEE International Conference on Data Mining*, pages 773–778, 2014. doi: 10.1109/ICDM.2014.82.

Maeve Duggan. Pew Research Center, "Online Harassment". http://www.pewinternet.org/2014/10/22/online-harassment/, October 2014. Accessed: 2020-10-25.

Hatebase. https://hatebase.org/about. Accessed: 2020-10-21.

Helsinki-NLP. Xed. https://github.com/Helsinki-NLP/XED, 2021. Accessed: 2021-01-15.

Pedro Rangel Henriques, Cristiana Araújo, Isabel Ermida, and Idalete Dias. Scraping News Sites and Social Networks for Prejudice Term Analysis. In Hans Weghorn and Luís Rodrigues, editors, *Proceedings of the 16th International Conference on APPLIED COMPUTING 2019*, pages 179–189, Cagliari, Italy, Nov 2019. ISBN 978-989-8533-95-1. doi: https://doi.org/10.33965/ac2019_201912L022.

Brian Heredia, Joseph D Prusa, and Taghi M Khoshgoftaar. Location-based twitter sentiment analysis for predicting the us 2016 presidential election. In *The Thirty-First International Flairs Conference*, 2018.

Věra Jourová. How the code of conduct helped countering illegal hate speech online. https://ec.europa.eu/info/sites/info/files/hatespeech_infographic3_web.pdf. Accessed: 2020-09-26.

Ana Filipa Vilela Pereira. SAQL: Query Language for Corpora with morpho-syntactic annotation. Master's thesis, Minho University, Braga, Portugal, March 2022. MSc dissertation.

Peter Norvig. http://norvig.com/spell-correct.html. Accessed: 2020-11-03.

Robert Plutchik. A psychoevolutionary theory of emotions. *Social Science Information*, 21 (4-5):529–553, 1982. doi: 10.1177/053901882021004003. URL https://doi.org/10.1177/053901882021004003.

Rui Rodrigues, Cristiana Araújo, and Pedro Rangel Henriques. NetLangEd, A Web Editor To Support Online Comment Annotation. In Ricardo Queirós, Mário Pinto, Alberto Simões, Filipe Portela, and Maria João Pereira, editors, *10th Symposium on Languages, Applications and Technologies (SLATE 2021)*, volume 94 of *Open Access Series in Informatics (OASIcs)*, pages 15:1–15:16, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-202-0. doi: 10.4230/OASIcs.SLATE.2021.15. URL https://drops.dagstuhl.de/opus/volltexte/2021/14432.

Rui Pedro Barbosa Rodrigues. NetLangEd, an editor to support Comment Analysis. Master's thesis, Minho University, Braga, Portugal, April 2022. MSc dissertation.

Pawel Sobkowicz, Michael Kaschesky, and Guillaume Bouchard. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, 29(4):470 – 479, 2012. ISSN 0740-624X. doi: https://doi.org/10.1016/j.giq.2012.06.005. URL http://www.sciencedirect.com/science/article/pii/S0740624X12000901. Social Media in Government - Selections from the 12th Annual International Conference on Digital Government Research (dg.o2011).

C. B. Stone and Q. Wang. From Conversations to Digital Communication: The Mnemonic Consequences of Consuming and Producing Information via Social Media. *Topics in Cognitive Science*, pages 774–793, 2019. ISSN 1756-8765.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, June 2011. ISSN 0891-2017. doi: 10.1162/COLI_a_00049. URL https://doi.org/10.1162/COLI_a_00049.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Stefanie Ullmann and Marcus Tomalin. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*, 2019.

Bertie Vidgen and Taha Yasseri. Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics*, pages 66–78, 2019.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P12-3020.

Wolfgarbe. Symspell. https://github.com/wolfgarbe/SymSpell, 2021. Accessed: 2021-01-15.

Ziqi Zhang. Hate Speech Detection: A Solved Problem?The Challenging Case of Long Tail on Twitter. *Semantic Web Journal*, 2018.