NetAC, An Automatic Classifier of online Hate Speech Comments

31st March 2021

Jorge Brandão Gonçalves
Constança Elias
Cristiana Araújo
Maria Araújo
Pedro Pinheiro
Pedro Rangel Henriques

Escola de Engenharia
Universidade do Minho

WORLD CIST2021
MARCH 30-31 TO APRIL 1-2, TERCEIRA ISLAND, AZORES, PORTUGAL

NetLang Project

- **Analysis** of the comments present in online newspaper sites and online social platforms for detecting **hate speech** .

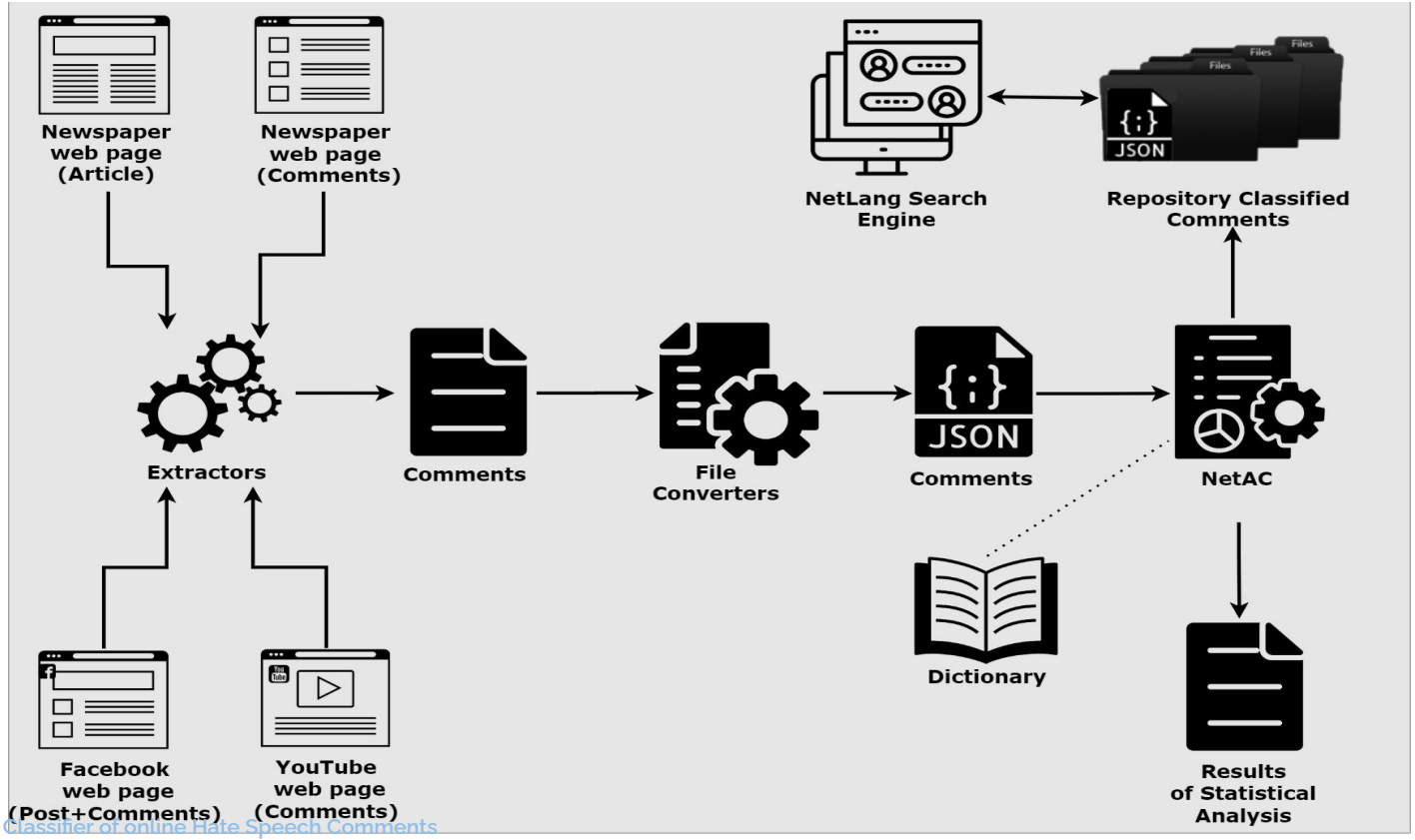
- Based on a comparable corpus of online texts in **Portuguese** and **English**.

NetLang Project

A Categorization Table based on **keywords** was built to assist the search for texts to be analysed by Linguists or Experts in Social Science areas.

TYPES OF PREJUDICE	SOCIOLINGUISTIC VARIABLES		KEYWORDS (English)	KEYWORDS (Portuguese)
	Hyperonym	Hyponym		
HOMOFOBIA	Sexual Identity	General	Homophobia, Gay, Queer, LGBT, Homophobic, Homoerotic, Homosexual	Homofobia, Gay, LGBT, homofóbico, Homoerótico, Homossexual)
		Female homossexuality	Lesbian, Battle-axe, Butch woman, Dike, Tomboy	Fufa, Machona, Maria-rapaz, Matrafona, Lambe-c*nas

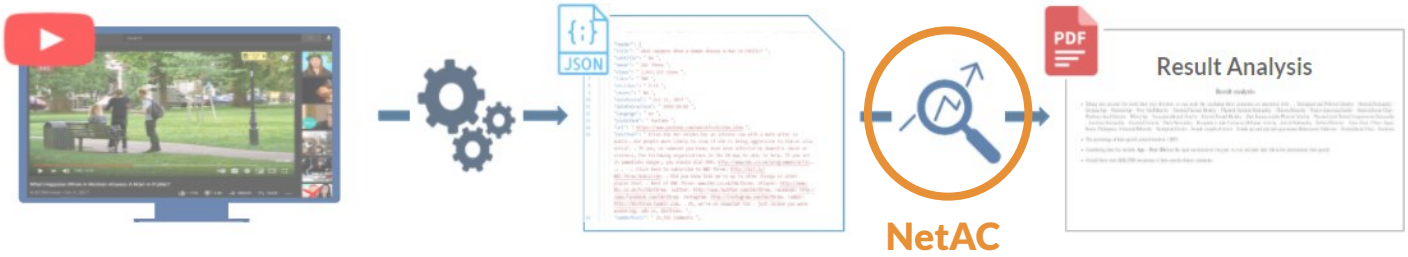
NetLang Architecture



Motivation for NetAC

➔ **Automate** the process of **searching** for the **frequency** of occurrence of *keywords* in each *category* based on a Categorization Table.

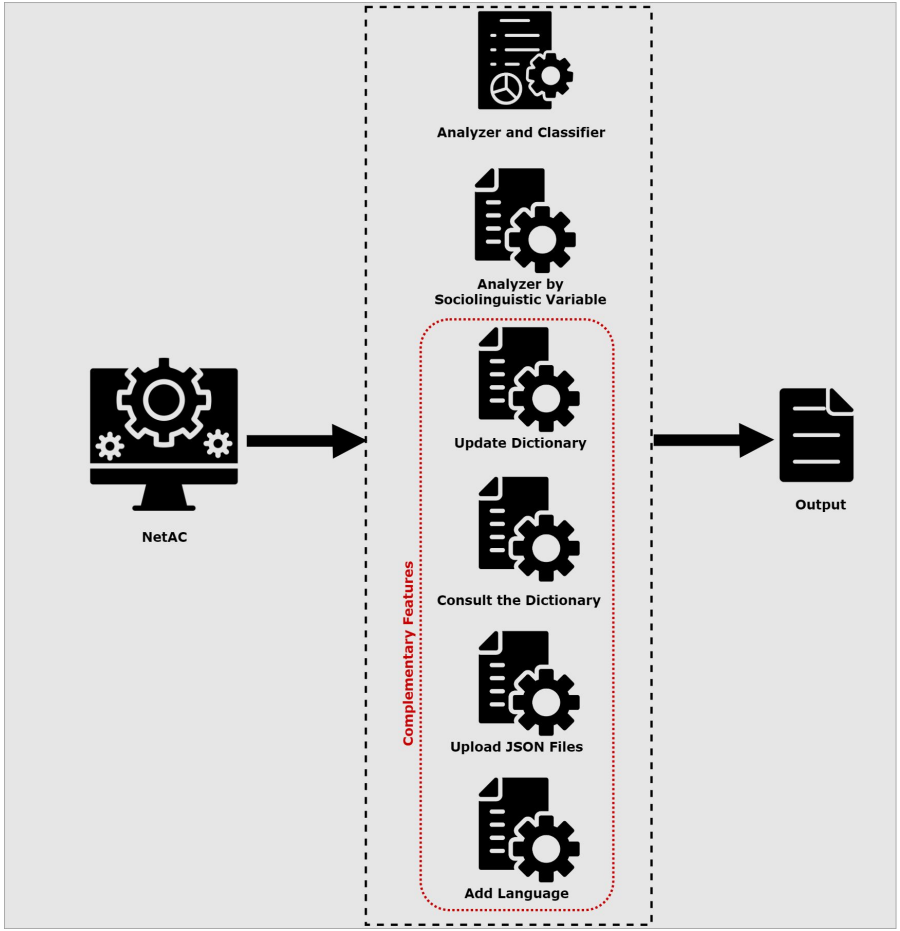
➔ Propose a **classification** for each **comment** and for the overall **text**.



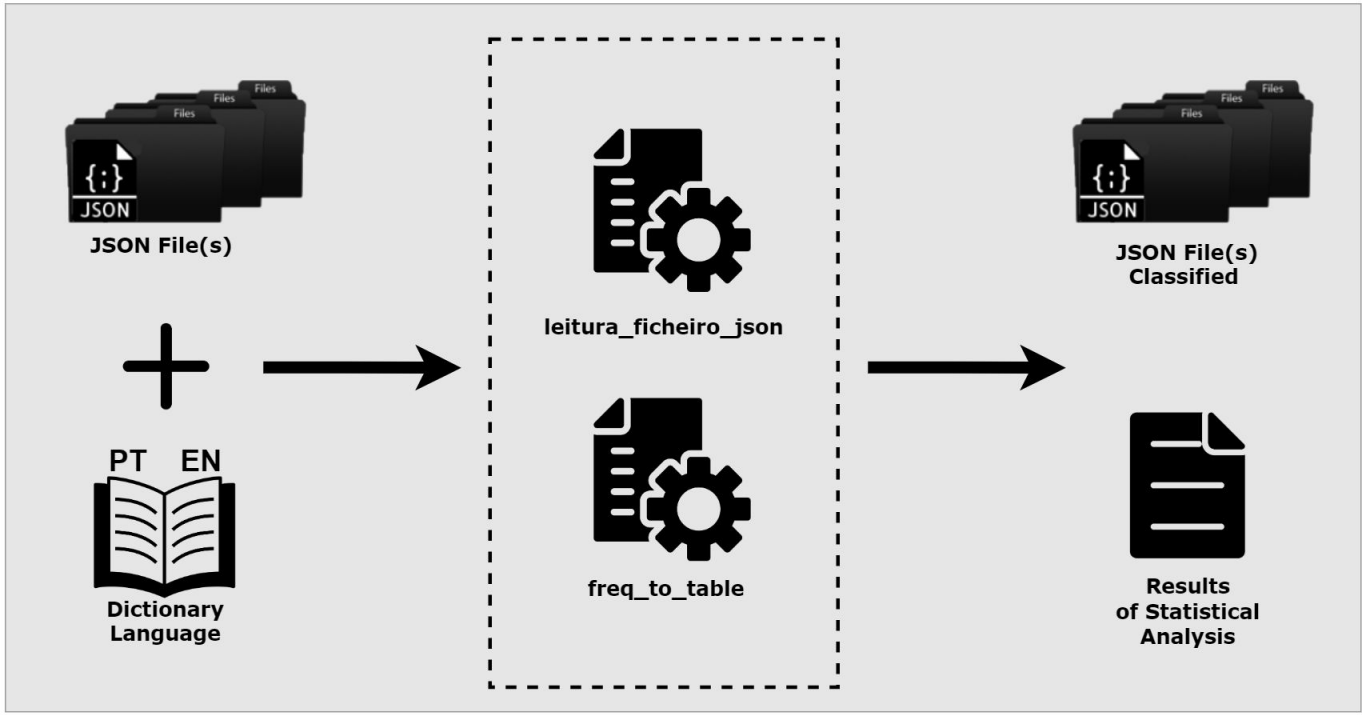
NetAC

NetLang Analyzer and Classifier

- **Main objective:** analyze and classify files (**Analyzer and Classifier**).
- Complementary features



NetAC Analyzer and Classifier



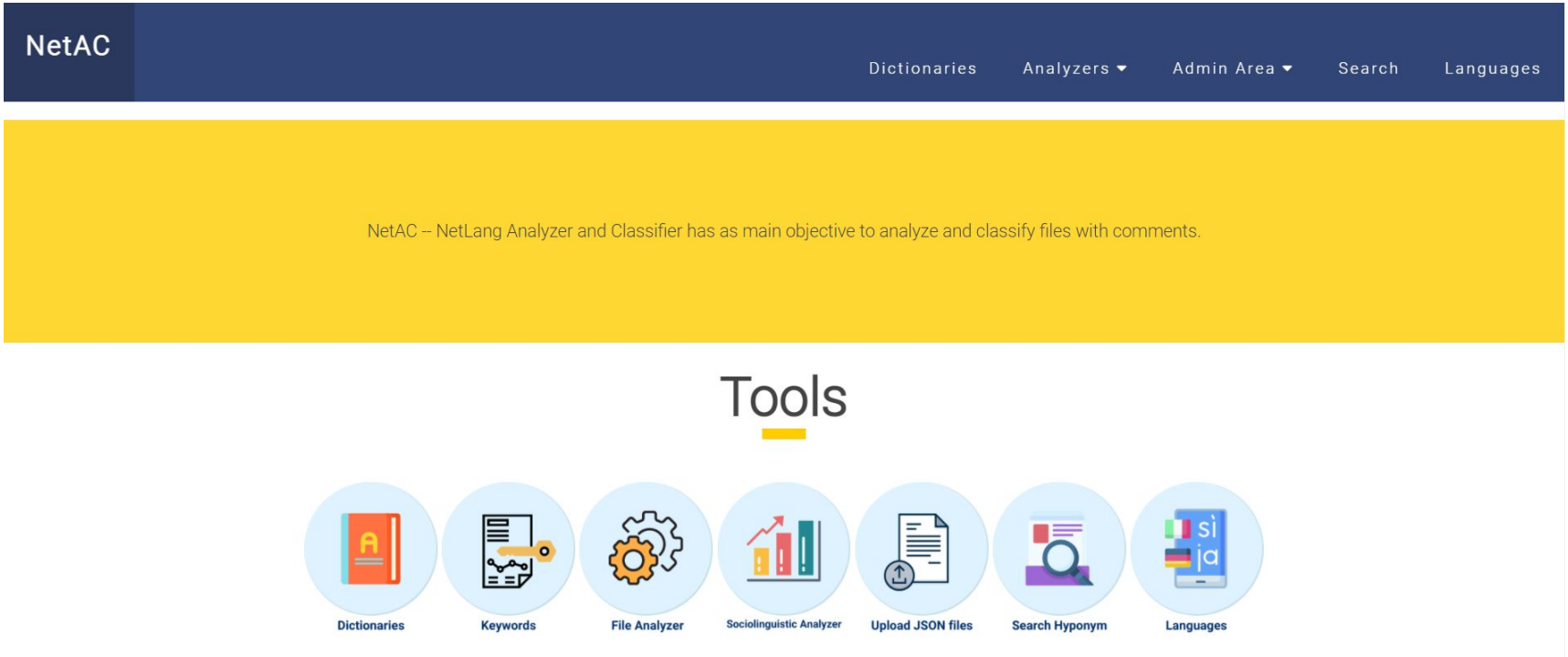
NetAC

Complementary Features

- Adding/ Removing New Keywords
- Dictionary Consulting
- Adding New Languages
- Uploading files

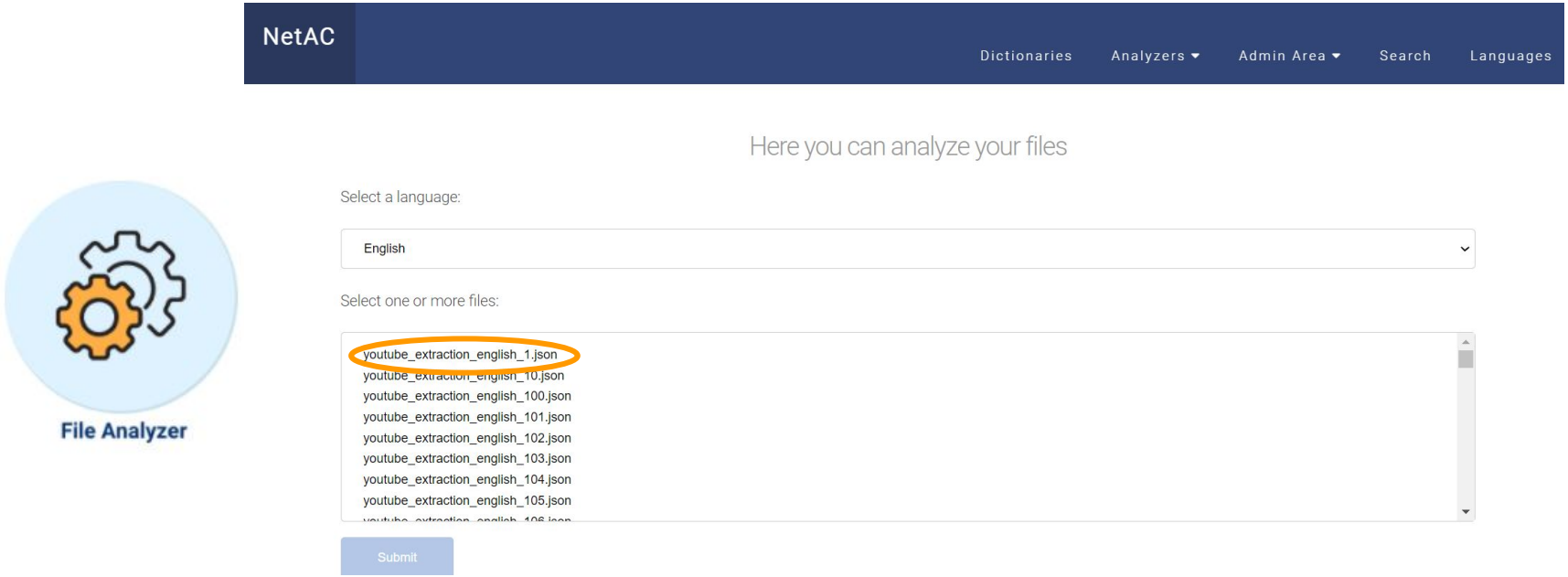
Web Application

http://netlang-corpus.ilch.uminho.pt:10100/



Case Study

Example of using the **File Analyzer** tool



Case Study

Results

1. General Summary

1 files were analyzed.

Overall there were **9026** occurences of hate speech related words in **557940**.

2. Detailed Analysis

Youtube_extraction_english_1

The predominant sociolinguistic variable in this file is **Gender - General**.

Case Study

Results

3. PDF files

This is a list of PDF files with the analyzis results of the files you just uploaded. Click on them to load/download the content.

For each file, two tables are shown.

A first analysis of hate speech by comments

the second makes a synthesis of the sociolinguistic variables found in the post and the associated hate speech.

At the end of the file, there is a summary of the analysis.

TabelaFreq_youtube_extraction_english_1.pdf

4. Classified JSON files to download

youtube_extraction_english_1.json

Case Study

Table 1: Summary of the results per comment

Comment	KeyWords	Sociolinguistic variables (Hiper - Hipo)	Hate Speech Frequency	Hate Speech Frequency(%)
If you guys believe that happen only in USA and Europe your so wrong here in Brazil a man was charged for the "crime of omission"because a women yelled and threw a glass of drink in a face of other man in this man react punching her, so the guy in the restaurant was in charge because he didn't defend this women that's **crazy** why some guy is obligated to defend some random women that he doesn't even know?	Crazy	Physical Identity - Physical (and Mental) Impairments	1/80	1.25
Look at what feminist have done. Made it so one **gender** is lower than other.	Gender	Gender - General	1/15	6.667
When this is taken to the next level. The women start using false rape accusation. And the whole society and media go on trial to destroy the man's life and self-respect. Always believe **woman** they sayno matter what man is always wrong and false. What a society we live in.	Woman	Gender - General	1/51	1.961
I read this story somewhere about a **woman** literally stealing a baby from her dad. The dad fought back and the **woman** started screaming that HE was trying to steal HER baby and all the people around them started helping her and beating him up while the **woman** almost escaped with the child. Even after the mom of the baby showed up and police were called and the situation cleared up [EMOJI] thank God! [EMOJI] the people who had come to help the thief [EMOJI] she got away [EMOJI] gave the dad this look like he was the villain of the story.Absolutely outrageous. To think that this assumption that the man has to be the perpetrator could have cost a family their future together	Woman	Gender - General	3/121	2.479
Because **woman** are helpless angels that everyone need to protect	Woman	Gender - General	1/10	10.0

Case Study

Table 2: Summary of the results per sociolinguistic variable

Sociolinguistic variables (Hiper - Hipo)	KeyWords	Number of occurrences	Frequency	Frequency(%)
Physical Identity - Physical (and Mental) Impairments	Crazy, Weird, Dumb, Mental, Blind, Psycho, Disabled, Handicapped, Lame, Lunatic, Deaf, Midget, Freak, Weirdo, Spastic	455	455/557940	0.08
Gender - General	Gender, Woman, Misogynist, Patriarchy, Sex, Dame, Sexism, Chick, Sexual, Misandry, Misogyny, Pussy pass, Chauvinist	5973	5973/557940	1.0699999999999998
Social Class - Working class	Common, Coarse, White Trash	118	118/557940	0.02
Age - Over 65s	Old, Elderly, Elder, Senior	207	207/557940	0.04
Ethnicity - Black	Black, African, Monkey, Buck, Crow, Colored, coloured, Ape	224	224/557940	0.04
Age - Youngsters	Nobody, Nothing, Zero, Brat	821	821/557940	0.15

Case Study

Result analysis:

- The percentage of hate speech related words is 1.6177.
- Considering that the variable **Gender - General** has the most occurences in the post, we can interpret that this is the predominant hate speech.
- Overall there were 9029/21765 occurences of hate speech related comments.

Conclusion

- We introduced a Web application that **analyses** and **classifies** posts and news articles into a predefined category of **Socially Inappropriate Discourse** (Hate or Aggressive speech).
- The results can be consulted in the PDF documents, accessed via web.
- This tool is being tested with the NetLang Corpus.
- The source language can be easily adapted to **other languages** (as long as they are based on the english alphabet) by creating a new dictionary for that languages keywords.

Conclusion

Future Work

- Design an experiment to be conducted with a Team of Linguists and Social Science Researchers for validating the automatic classification proposed by our tool.
- Design an experiment to assess the tool's usability.
- Use **artificial intelligence** approaches to refine the linguistic analysis and **overcome** problems like the **false positive** identification.