HATE SPEECH DETECTION

A case of generated data

THE DATA

ENGLISH

- Dynamically generated Hate Speech dataset
- Permutation of words in order for labels to be switched from Not Hateful to Hateful (Perturbation)
- Aim: robust algorithms for hate speech detection
- Binary Hate (53.8%) vs NotHate annotation
- “Target” annotation: 42 categories of hate

FRENCH

- French portion of a dataset for Multilingual and Multi-Aspect Hate Speech Analysis
- 4.014 entries

DYNAMICALLY GENERATE HATE SPEECH DATASET

First Step: Text Cleaning

Binary classification:
complete dataset
(41.144 entries)

Multiclass classifaction: six most
common hate categories
(7.610 entries)

BINARY CLASSIFICATION

Hate (0) vs NotHate (1)

MULTICLASS CLASSIFICATION

Women (5)
Black People (0)
Jewish People (2)
Muslim People (3)
Trans People (4)
Gay People (1)

THE METHODS

CLASSIFICATION PART 1

- Logistic Regression, XGBoost and SGD Classifier with Tf-Idf
- XGBoost and SGD Classifier with Word2Vec word embeddings trained on the training set
- XGBoost with W2V and FastText pretrained word embeddings
- All methods applied to both binary and multiclass classification

CLASSIFICATION PART 2

- Convolutional Neural Network with W2V and FastText pretrained word embeddings (9 Conv1D layers)
- Bidirectional Long Short Term Memory Neural Network with W2V and FastText pretrained word embeddings (6 Bi-LSTM layers)
- All methods applied to both binary and multiclass classification

THE METHODS

WORD2VEC TRAINING

- Context Window of 2 words
- Only using words that appear at least 10 times
- Vector size per word: 300, as for W2V and FastText provided by Gensim

THE METHODS

KEYWORDS AND KEYPHRASES

WordClouds, 50 words
KeyBert with one word, two words, and key phrases
BerTopic

BEFORE CLASSIFICATION

FOR ALL EXPERIMENTS

Stopwords removal

Exceptions: “they”, “them”, “no”, “not, “don’t”

Tokenization and Stemming (SnowballStemmer)

CLASSIFICATION PART 1: RESULTS

BINARY CLASSIFICATION

Algorithm - TF-IDF	Accuracy
Logistic Regression	0.69
XGBoost	0.67
SGD Classifier	0.68

MULTICLASS CLASSIFICATION

Algorithm- TF-IDF	Accuracy
Logistic Regression	0.87
XGBoost	0.87
SGD Classifier	0.90

CLASSIFICATION PART 1: RESULTS

BINARY CLASSIFICATION

Algorithm - Word Embeddings	Accuracy
XGBoost Trained W2V	0.59
SGD Trained W2V	0.59
XGBoost Pretrained W2V	0.61
XGBoost Pretrained FastText	0.61

MULTICLASS CLASSIFICATION

Algorithm - Word Embeddings	Accuracy
XGBoost Trained W2V	0.64
SGD Trained W2V	0.31
XGBoost Pretrained W2V	0.78
XGBoost Pretrained FastText	0.80

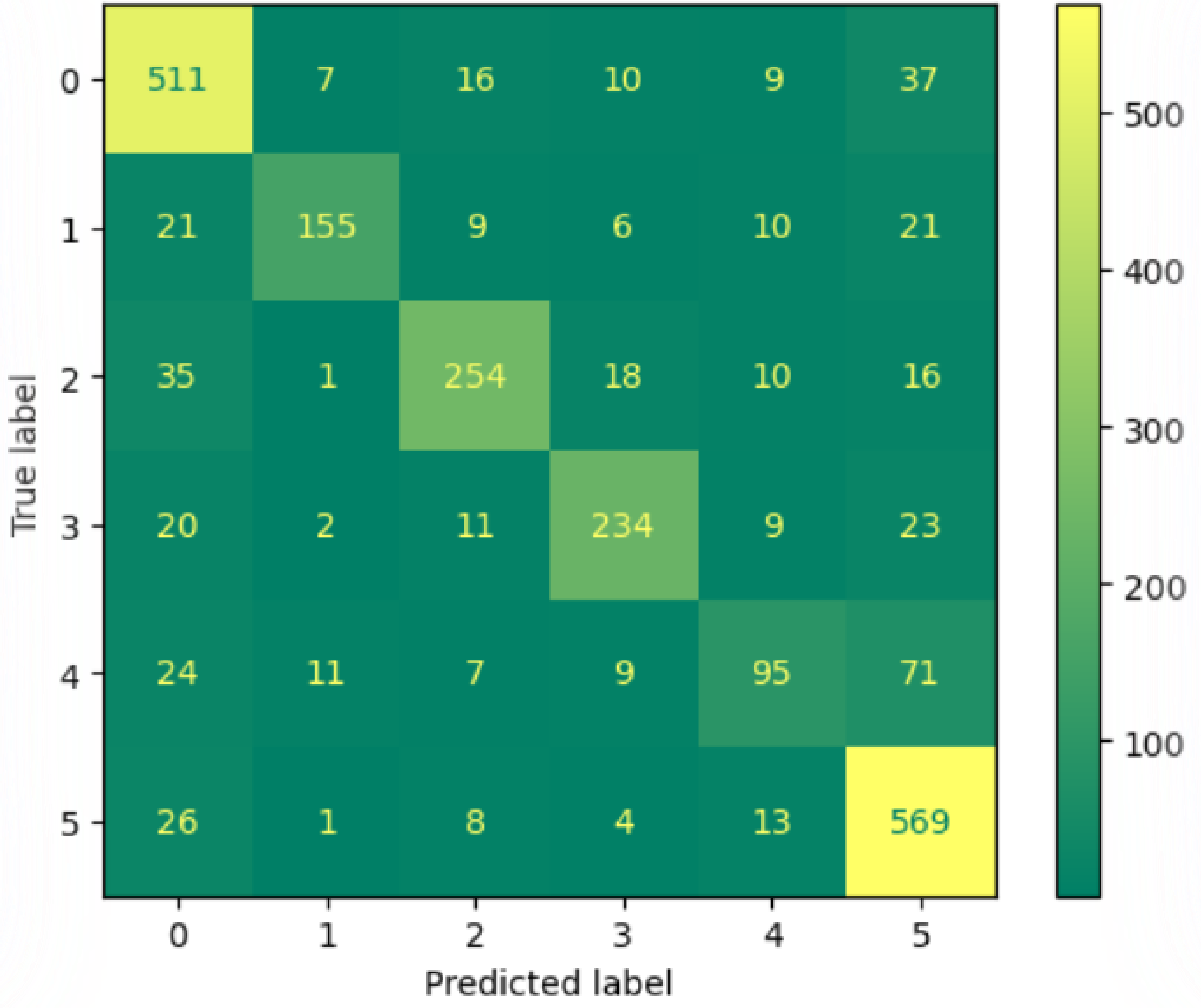
CLASSIFICATION PART 1: RESULTS

BINARY CLASSIFICATION

- For Tf-Idf, all values of precision and recall are between 64% and 72%
- When training the embeddings SGD obtains a 92% recall on the Hate category, but 57% precision and a 19% recall for the NotHate category: the algorithm tends to classify most text as hateful
- Not enough data and too short a window to train embeddings, similar words for both categories
- Pretrained word embeddings do not significantly improve results
- Need for methods which better encapsulate context

MULTICLASS CLASSIFICATION

- For Tf-Idf, more specific categories - with a more specific vocabularies lead to improved results; with SGD values for precisions or recall are over 80% for all categories
- Trained W2V doesn't improve results, with SGD completely misclassifying 3 of the 6 categories
- Pretrained word embeddings do not significantly improve results
- In all models but one the lower values for recall are detected for category 4



XGBoost Confusion Matrix for Multiclassification with FastText. For all models but one, recall for category 4 is the lowest; entries are mostly misclassified as category 5.

As category 4 is the “Trans” category and 5 is “Women”, this could be because of the similarity in the vocabulary for the two; hate speech against trans people often refers to terms relating to gender, such as “Woman”, “Men” or “Girl”.

CLASSIFICATION PART 2: RESULTS

BINARY CLASSIFICATION

Algorithm	Accuracy
CNN W2V	0.6775
CNN FastText	0.6812
Bi-LSTM W2V	0.6952
Bi-LSTM FastText	0.6923

MULTICLASS CLASSIFICATION

Algorithm	Accuracy
CNN W2V	0.7587
CNN FastText	0.7652
Bi-LSTM W2V	0.8178
Bi-LSTM FastText	0.8296

CLASSIFICATION PART 2: RESULTS

BINARY CLASSIFICATION

- Results for accuracy on pair with the best performance obtained by XGBoost with TF-IDF
- Overfitting: training stops as early as after 12 epochs
- Values of binary cross-entropy all higher than 0.56
- Improvement of recall values for Hate category
- Need for more complex (deeper) network structure to avoid overfitting

MULTICLASS CLASSIFICATION

- Results for accuracy lower than those obtained with TF-IDF
- Overfitting: training stops as early as after 16 epochs
- Values of categorical cross-entropy all higher than 0.54 and as high as 0.89
- Values of accuracy and recall lower or comparable than those of Part 1
- Need for more complex (deeper) network structure to avoid overfitting

KEYWORDS AND KEYPHRASES: RESULTS

ENGLISH vs FRENCH

- WordClouds for Hate, NotHate and Specific Categories
- KeyBert for Hate, NotHate and the six hate categories
- BerTopic with documentation suggested pipeline

- KeyBert for “individual”, “other” ,“arabs” and“african descent” categories
- BerTopic with documentation suggested pipeline

For both languages, KeyBert for key phrases and BerTopic use TKeyphrase TfIdf Vectorizer,

DATA PREPARATION

ENGLISH

- WordClouds: stopwords removal
- BerTopic: automatically removes stopwords

FRENCH

- Text Cleaning
- BerTopic: automatically removes stopwords

DISCLAIMER: THE FOLLOWING SLIDES CONTAIN UNCENSORED DEROGATORY SPEECH AND SLURS, OFTEN VIOLENT IN NATURE.

NO HATE CATEGORY CLOUD

“People”, “women”, and “don’t” are three of the most common terms for both clouds; ulteriorly, the stopwords could have been expanded with terms such as “re”.

HATE CATEGORY CLOUD

WordClouds showed that both the Hate and NotHate category contain generic and neutral terms, category names, other than several overlapping terms.

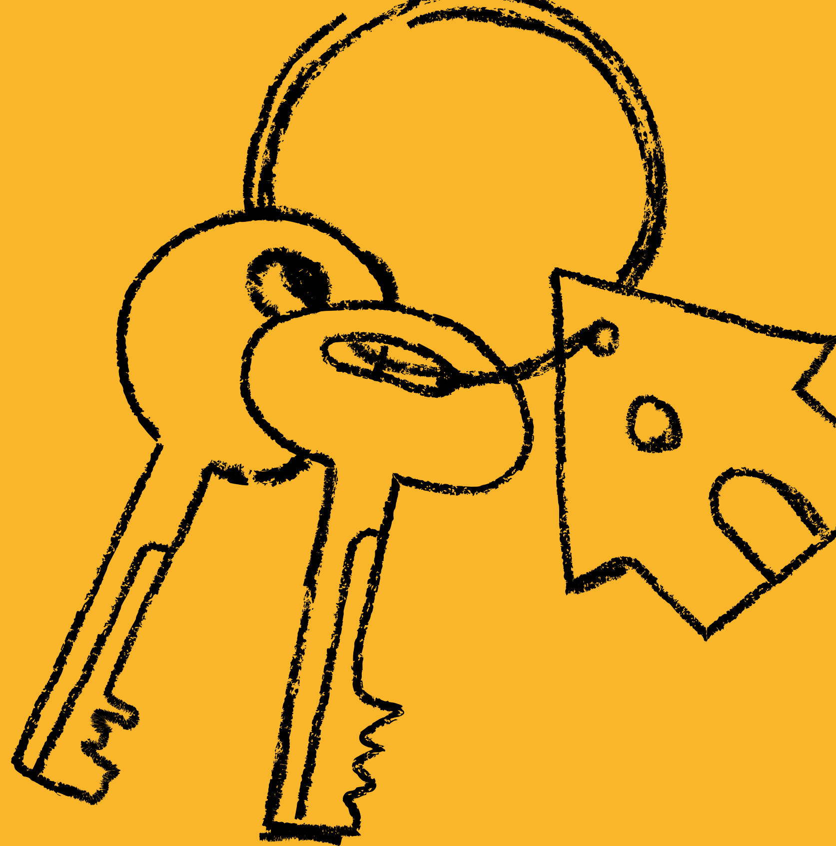
TRANS CATEGORY CLOUD

Even specific categories can have overlapping vocabularies: hate speech for trans people is strongly related to the "women categories"

JEWISH CATEGORY CLOUD

Slurs and insults are common, but also neutral terms such as "control" or "media": sentiment polarity is not easy to integrate in hate speech detection

KEYBERT: ENGLISH



"Immigrants dalits"

"Racists"

"Refugees"

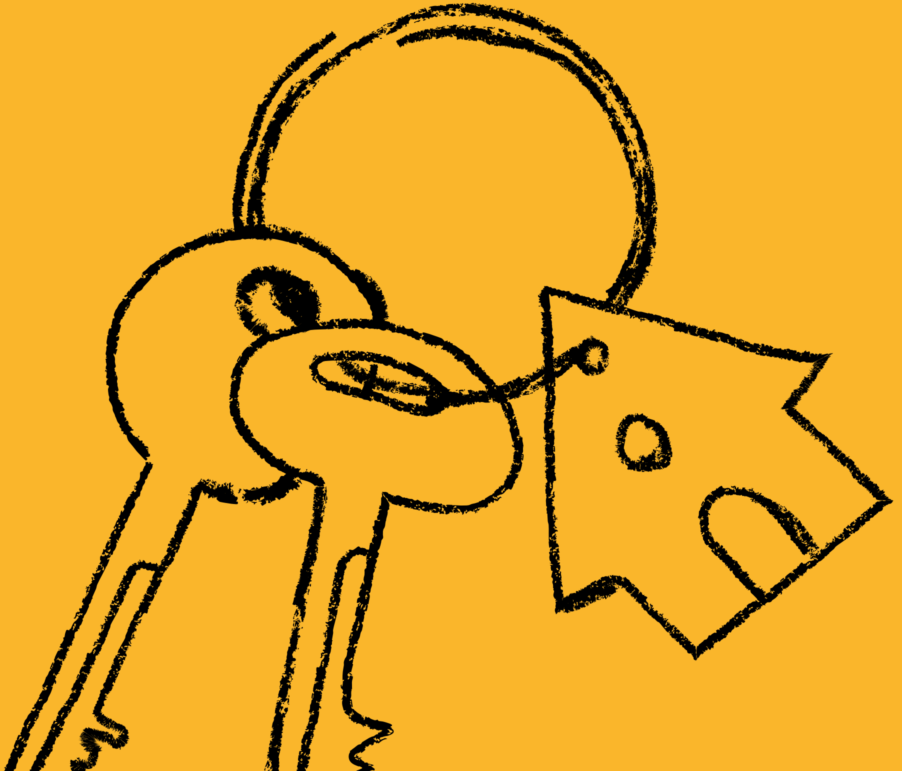
HATE

NOT HATE

"Pakistanis"

"Cunts"

"Foreigners Disgusting"



KEYBERT: ENGLISH

HATE SUBCATEGORIES

"Feminist"

"Patriarchy"

"Islamophobia"

Words with clear and strict relationship with the topic

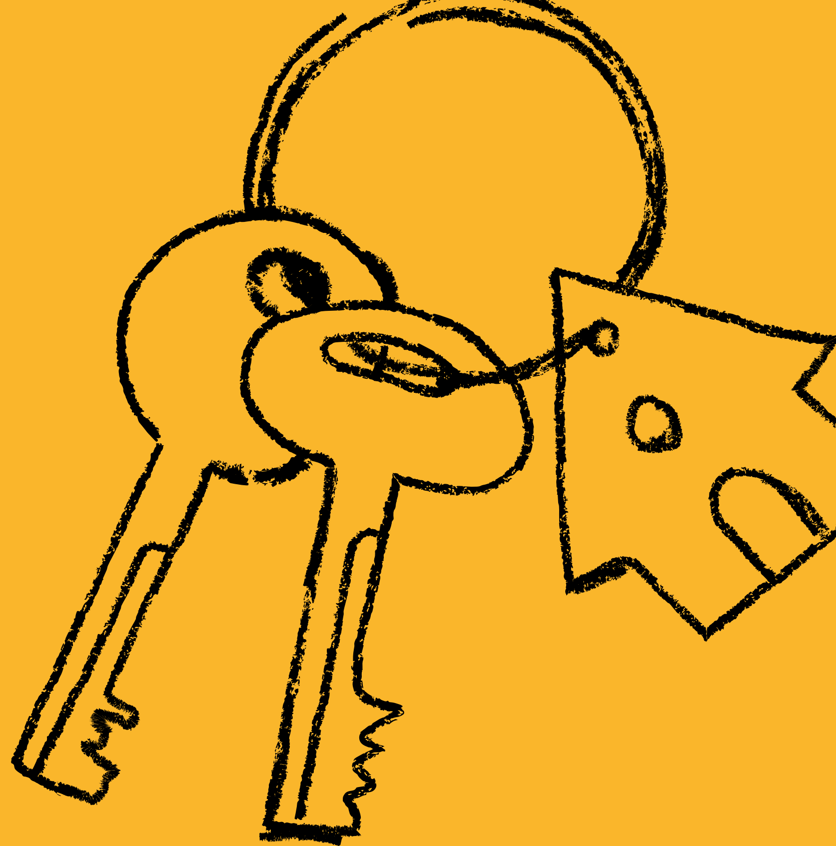
Insults, slurs, words with modified spelling, word + derogatory insult combinations

"Feminist bitch"

"Jewsssss"

"Trans freaks"

KEYBERT: FRENCH



"Mongole"

"Communiste"

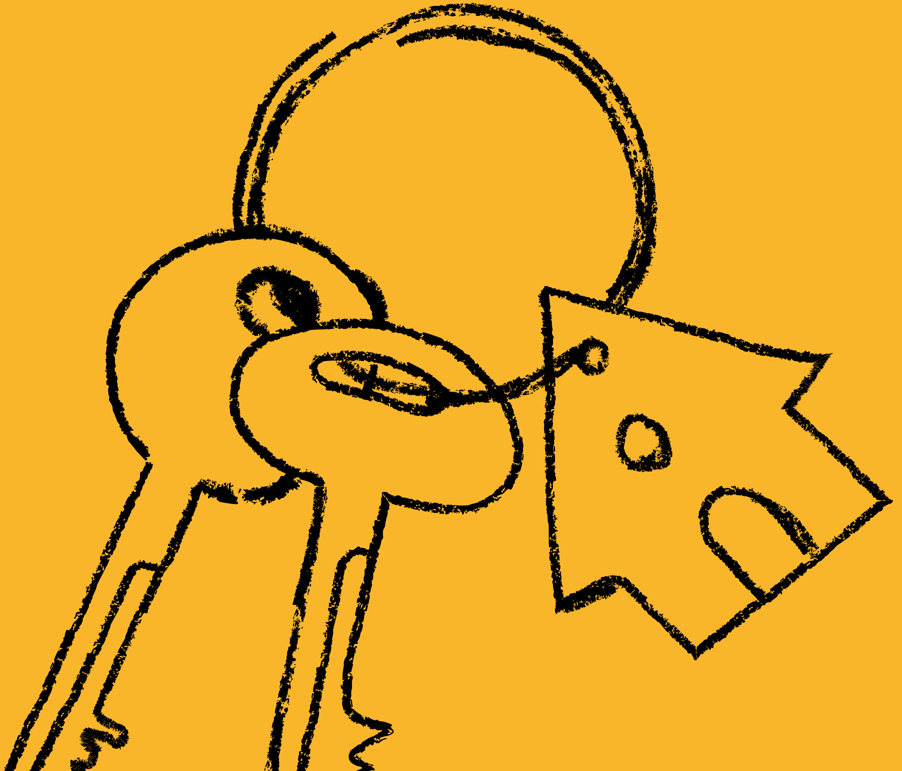
"Arabe"

Insults and slurs are similarly detected; the dataset seem to contain more political references and ableist insults

"Vraiment mongol"

"Militantiste"

"Macron terrorisme"



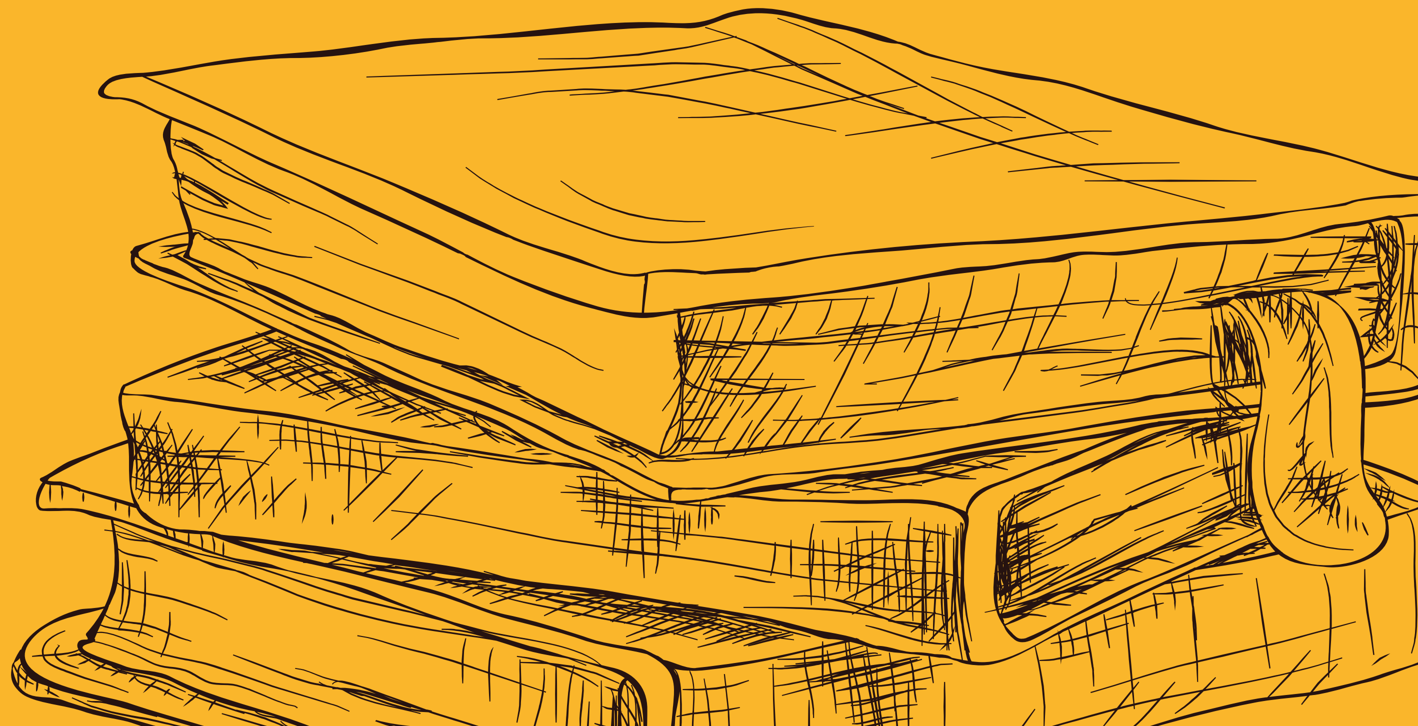
FRENCH VS ENGLISH

Beside the presence of slurs and derogatory terms, we notice some detected key phrases have almost direct correspondence in the two languages.

"Sale arabe" —*Translates to*→ *"Dirty arab"*

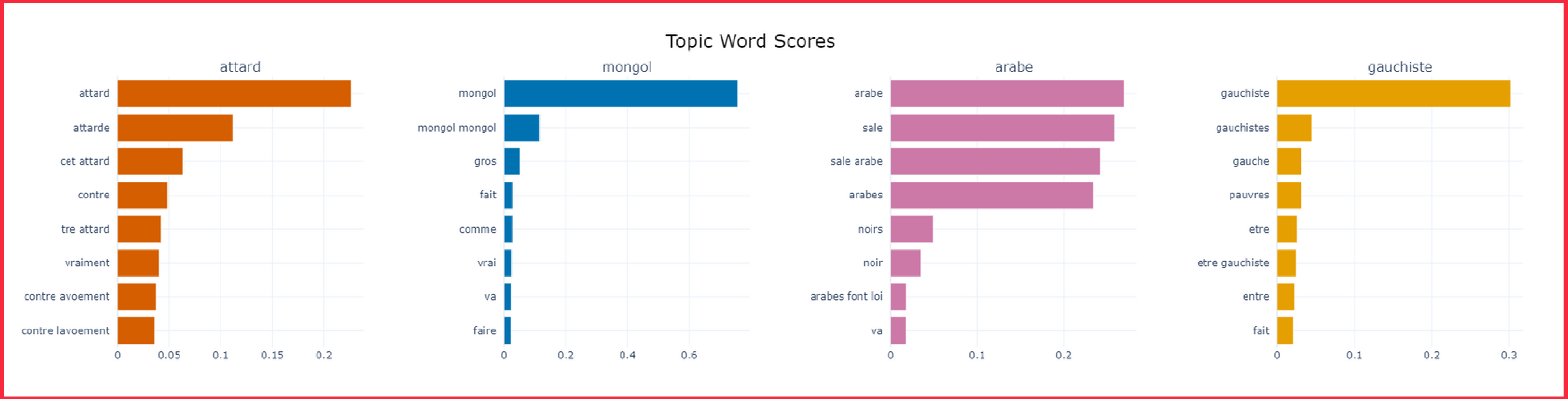
Similar in meaning to

"Dirty Muslim"



BERTOPIC: ENGLISH

- 14 Topics detected, with presence of similar ones that could be integrated
- One category of numbers, mispellings and social media terminology (“chad”)
- Coherence with previous findings and dataset category annotation
- Keywords connected to stereotypes: “Chinese” category contains “covid”, “Indians” contains “curry”
- Keywords connected to socio-political discussions: “officers” and ”white” connected to Black people



BERTOPIC:FRENCH

- 10 Topics are detected, but it shows the need for better text cleaning, specific to the French language: many of the detected words are generic, verbs or connectives
- As for English, some of the topics detected are duplicates, while the relevant keywords contain slurs and derogatory terms

Conclusions

1. While Binary Classification can achieve 69% accuracy, Multiclass Classification reaches much better results (up to 90% accuracy)

2. This can be explained by the much more specific vocabulary of hate subcategories, while the Hate and NotHate category present many overlapping common terms

3. The presence of common terms with neutral meaning makes it difficult to integrate sentiment polarity in classification; more adavanced techniques to leverage context and semanthics should be implemented

4. A first step to improve this work could be taken by restrciting the vocabulary for Binary Classification by removing common terms shared by the Hate and NotHate categories

THANK YOU FOR YOUR ATTENTION!

TEXT MINING AND SENTIMENT ANALYSIS
A.A. 2022/2023

MARTA CAMPAGNOLI

EMAIL ADDRESS:

marta.campagnoli@studenti.unimi.it

I.D. 928635