M-Phasis: A Feature-Based Corpus of Hate Online

Natural Language Processing

FEUP 2022/2023

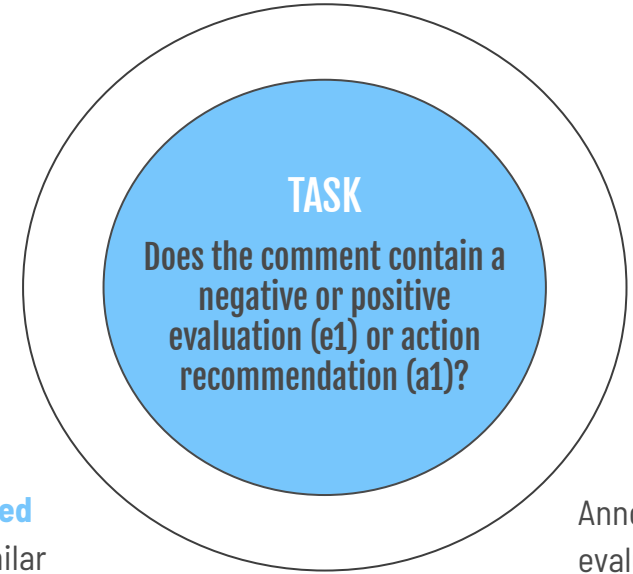
GROUP D

Maria Carneiro - up201907726@edu.fe.up.pt
Pedro Silva - up201907523@edu.fe.up.pt
Rodrigo Andrade - up201904967@edu.fe.up.pt

Domain Description

Corpus of ~ 9k **German** and **French** user **comments** collected from **migration-related news articles**.

Hate Speech Online analyzed by 4 annotators per language, with 23 features regarding 3 different tasks.



Comments collected from migration based articles from **January 2020** to **May 2020**, with at least 5 comments per article.

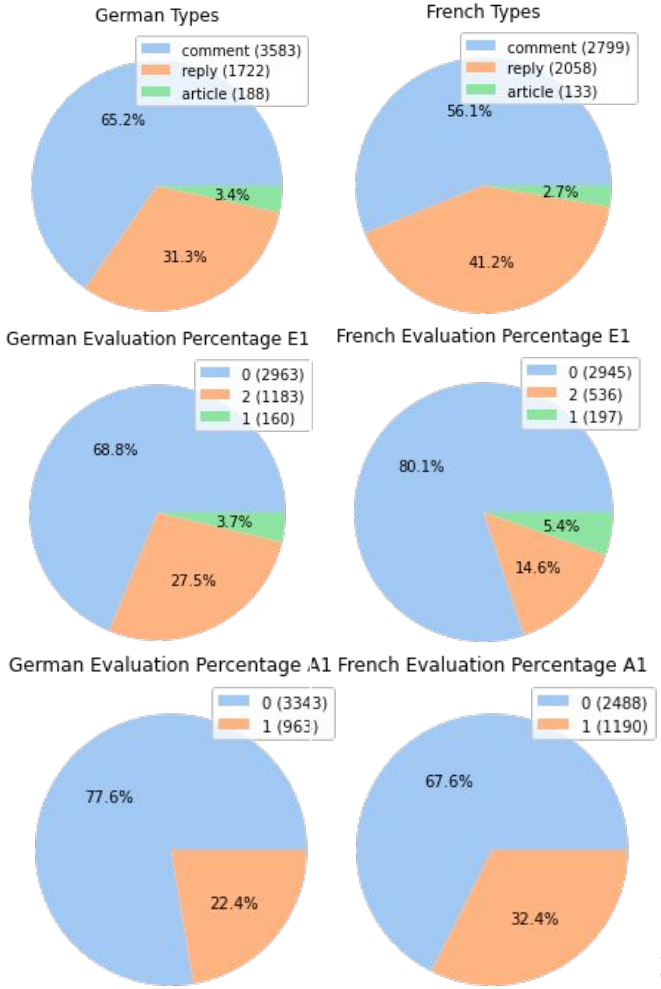
Comments not based on slurs lists to capture **implicit** forms of hate, derived from a diverse set of mainstream and fringe media outlets, collected in **threads** for **context-sensitive** analysis.

Refined original dataset into a **task-based** dataset, clustering annotations with similar meanings. Focused on evaluating the comment or it's action recommendation.

Annotated across 5 different modules: Negative evaluation, Positive Evaluation, Recommendation of an Action, Contrasts and Emotions, with regards to **explicit** and **implicit** meanings.

Exploratory Data Analysis

- **3678** comments for **French** across 5 news outlets (mostly '*figaro*' w/ 1007) and **4306** for **German**, across 6 news outlets (mostly '*tagesschau*' w/ 960).
- There are only a few articles from where the comments and replies were retrieved, since they are most of the content in the original dataset.
- Comments with positive evaluations were the ones with most words.
- Most evaluations were **positive** and **no action** was mostly recommended.



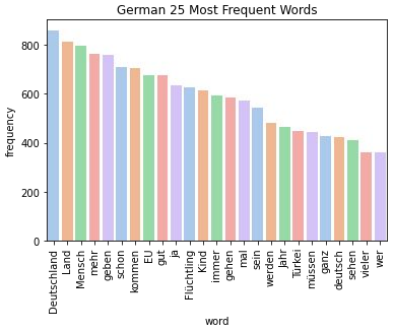
E1: 0 - Positive 1 - Negative and Explicit 2 - Negative and Implicit

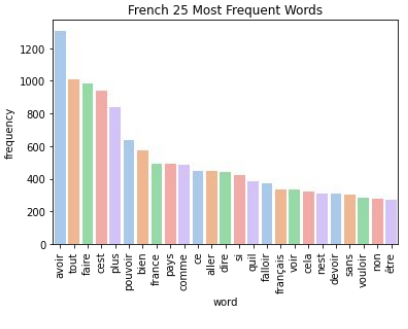
A1: 0 - No action 1 - Explicit Action

Exploratory Data Analysis

Word Frequency

- Some of the most frequent words in both languages reference migration in the EU context like 'Flüchtling' (refugee) or 'pays' (country).



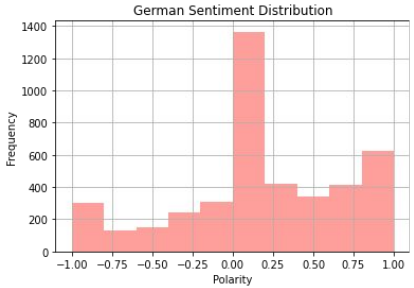


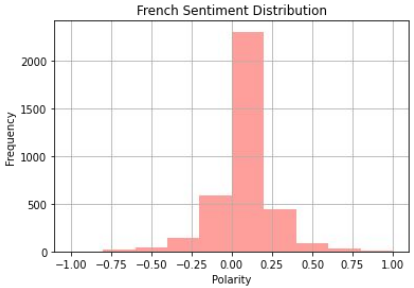
TF-IDF

- Most common words are generally the most important to a comment.

Sentiment Analysis

- Done by using TextBlob for both languages.
- Most comments for both languages had neutral sentiments.
- German had more polarity in their comments than French.





Data Preprocessing

Emoji Substitution

- Replacement of emoji by their textual form in each language.
- Emojis hold meaningful information in the context of an online comment that would otherwise be lost if removed.

Content Cleaning

- Removal of-non alphabetic chars from each comment, for both languages.

Tokenizing and Stemming

- Application of a stemmer for each language.
- Tokenization of the stemmed output, in order to build the corpus.
- Common stopwords hold meaning in comments and may affect their sentiment, so they were not removed.

Before

German

French

German

After

French

Models and Metrics

Metrics

- **Macro F1** is the main metric for evaluation.
- Dataset is **imbalanced**.
- Used in the original paper [1], so we adopt it as well to have a means of comparison.

Models

- **Hyperparameter tuning** + **Cross validation with 10 folds** (GridSearchCV)
- **Smote sampling**
- LogisticRegression
  - L2 penalty; Lbfgs and liblinear solvers; C = 0.1, 1, 10
- SVC (support vector machine)
  - Linear and rbf kernels; scale and automatic kernel coefficient; C = 0.1, 1, 10
- MultinomialNB (Naive Bayes)
- DecisionTree
  - Gini and entropy criterion; best and random splitter

Sparse Representations

Bag of Words

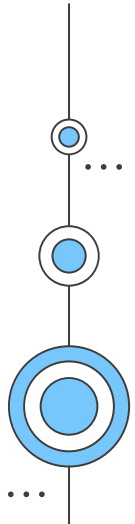
- Order of words is disregarded.
- Texts are represented by a multiset of its words.

N-grams

- In addition of representing the text as a collection of words (unigram) we also considered them as groupings of two words (bigrams).
- Some order is preserved with this representation.

Count and TF-IDF

- Features can be extracted.
- Count : the number of times each word appears in the text.
- TF-IDF : how important a word is to a document in the corpus.



Sparse Representations – Results

Method	Features	F1 – measure (FR – A1)	F1-measure (DE – A1)	F1 – measure (FR – E1)	F1 – measure (DE – E1)
CountVectorizer - Unigrams	12 073 (fr) 22 102 (de)	0.5469	0.5426	0.4710	**0.4821**
Tfldf - Unigrams	12 073 (fr) 22 102 (de)	**0.6119**	**0.6874**	**0.5366**	0.4655
Tfldf - Bigrams	30 000*	0.5562	0.4616	0.3675	0.3618

* Feature selection based on the chi2

Dense Representations

Word Embeddings

- Word Embeddings for each language from "ELMo for many langs" [2][3].
- Word vectors are learned functions of the internal states of a deep bidirectional language model [4], trained on a large corpus from the Web.
- Vector size is 1024.

Aggregation Methods

- Concatenation of word embeddings.
- The length of the sequence was 55 for German dataset (mean of 58.207 words per sentence).
- The length of the sequence was 40 for French dataset (mean of 41.4363 words per sentence).
- Mean of embeddings by component
- Concatenation of mean of embeddings by axis and standard deviation by component.

Performance

- Contrary to expectations word embeddings did not significantly increase performance.
- Explained by the nature of the documents used to train embeddings: Wikipedia articles and other web documents vs User comments.
- Less choice of pre-trained embeddings in German, French when compared to English.

Dense Representations – Results

Aggregation Function	Features	F – measure (FR – A1)	F-measure (DE – A1)	F – measure (FR – E1)	F – measure (DE – E1)
Concatenation	30000*	0.5342	0.5831	0.5048	0.4126
Mean	1024	0.5010	**0.6480**	**0.5522**	**0.4725**
Mean + Std	2048	**0.6285**	0.6320	0.5160	0.4381

* Feature selection was used based on the chi2

Comparison with State-of-the-Art

Feature	F - measure (FR - A1)	F - measure (DE - A1)	F - measure (FR - E1)	F - measure (DE - E1)
M-Phasis [1]	**0.669**	**0.723**	**0.596**	**0.556**
Our best model	0.629	0.687	0.552	0.482

Error Analysis – German Dataset

Most common Mistakes

- The classes the model confused the most were 0 (positive evaluation) with 2 (implicit negative) - 86%.
- The model does not over-classify as negatives (44%) or positives (50%).

Subjectivity and Lack of context

Sentence	Predicted	Expected
"Wow, dear Mr. Aust, there is nothing to add, you are absolutely right!"	0	1
"Quite simply, the right to asylum must be suspended or excluded in these cases. Otherwise, this proposal really makes no sense, you recognized that correctly."	0	2

Error Analysis – German Dataset II

Countries and People

- In 57 % of the cases when a text is classified incorrectly as a negative explicit evaluation (1), a country or a group of people is mentioned.
- The model also over-blacklists when people are mentioned.

Sentence	Predicted	Expected
"As a people, Poland, the Czech Republic and Hungary claim their fundamental right ..."	1	2
"Yes, exactly. It would be best to tell the people from Africa that Germany ..."	1	2
"Oh yes? 75% of Poles find Ukrainians unsympathetic ..."	1	2
"There was probably trouble Mr. Hackensberger? Your last report was much more objective..."	1	0
"that's what they were told. Just hoping for the savior Merkel, who is revered..."	2	0

Error Analysis – French Dataset

Most common Mistakes

- Just like in German, the classes the model confused the most were 0 (positive evaluation) with 2 (implicit negative) - 47%. The number is significatively lower than the German, probably due to the nature of the news outlets (more fringe in French).
- Overall less over-blacklisting than German

Subjectivity and Lack of context

Sentence	Predicted	Expected
"He wasn't just a whistleblower!!!"	2	0
“they are young and generally in good health, so they risk NOTHING…”	1	2

Main Insights

- **Results** were **similar but slightly lower** in comparison with results from the original report. That could be due to the limitation in using deep-learning models, which were implemented by M-Phasis, so we had **overall good results**.
- Very **subjective dataset** with a **topical bias** towards migration. Context that was given to annotators wasn't given to the model, which could've helped to distinguish positive or negative explicit classifications from negative implicit ones.
- **Lack of word embeddings** for both German and French, in comparison with English. The ones that did exist were not pre-trained with a similar domain to the data which could've helped in achieving better results.

Questions?

References

- [1] - Dana Ruiter, Liane Reiners, Ashwin Geet D'Sa, Thomas Kleinbauer, Dominique Fohr, Irina Illina, Dietrich Klakow, Christian Schemer, and Angeliki Monnier. 2022. Placing M-Phasis on the Plurality of Hate: A Feature-Based Corpus of Hate Online. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 791–804, Marseille, France. European Language Resources Association.
- [2] - Che, Wanxiang, et al. "Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation." arXiv preprint arXiv:1807.03121 (2018).
- [3] - Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In Proceedings of the 21st Nordic Conference on Computational Linguistics, pages 271–276, Gothenburg, Sweden. Association for Computational Linguistics.
- [4] - Sarzynska-Wawer, Justyna, et al. "Detecting formal thought disorder by deep contextualized word representations." Psychiatry Research 304 (2021): 114135.