HATE SPEECH CLASSIFICATION

TEXT MINING AND SEARCH COURSE

CARLO ARPINI 918543 - EMMANUELE LOTANO 918608 - CHIARA MARIANI 918354

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA, CDLM DATA SCIENCE

MAIN STEPS

PREPROCESSING

This involves cleaning and standardizing tweets by removing irrelevant elements, handling language inconsistencies and preserving meaningful information

TWEETS REPRESENTATION

We use BoW for frequency-based representation, GloVe for capturing semantic relationships through word embeddings, and sBERT for generating sentence-level contextual embeddings

TWEETS CLASSIFICATION

We use XGBoost, MLP and SVM as classifiers to assign tweets to specific cyberbullying types, combining all three of them with each tweets representation

CLUSTERING AND TOPIC MODELING

We use the k-means algorithm and only on the best result, obtained with sBERT, we perform topic modeling on each cluster to compare them with the ground truth classification of our data

PREPROCESSING

- Remove links, mentions, emoticons made with letters and re-tweet
- Remove hashtag symbols and store each hashtag in a dedicated column
- Substitute html tags

- Deal with the language problem: use the langid library then correct false negatives and false positives
- Substitute emojis and currencies into literal words
- Remove numbers, the remaining punctuations, white spaces and empty tweets

- Transform tweets into lower case
- No remove stopwords
- No perform lemmatization

03.

TWEETS REPRESENTATION

BoW

Each unique word in each tweet is represented as a feature, the value of each feature is the count of its occurrences in the tweet

Implemented using Count Vectorizer without stop words

GloVe

It generates static word embeddings, where each word has a single, context-independent vector based on co-occurrence statistics

Implemented using a pre-loaded word embeddings file from Github

sBERT

It produces contextualized embeddings taking into account the surrounding context of a word within each tweet

Implemented using a pre-loaded contextual word embedding

TWEETS CLASSIFICATION

Our three classifiers: *XGBoost, Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM)*

What have we obtained?

Compare the accuracy values to extract insights about the performance of different document representation

	XGB	MLP	SVM
BoW	0.834	0.785	0.824
GloVe	0.768	0.774	0.796
sBERT	0.792	0.788	0.830

Analyze the classification report to extract insights on how each model classifies different classes of tweets

```
XGBoost Accuracy: 0.834597072826662
XGBoost Classification Report:
              precision    recall  f1-score   support

           0       0.63      0.41      0.50      1723
           1       0.99      0.97      0.98      2012
           2       0.98      0.97      0.98      1899
           3       0.96      0.94      0.95      1974
           4       0.91      0.83      0.87      1918
           5       0.56      0.83      0.67      1816

    accuracy                           0.83     11342
   macro avg       0.84      0.83      0.82     11342
weighted avg       0.85      0.83      0.83     11342
```

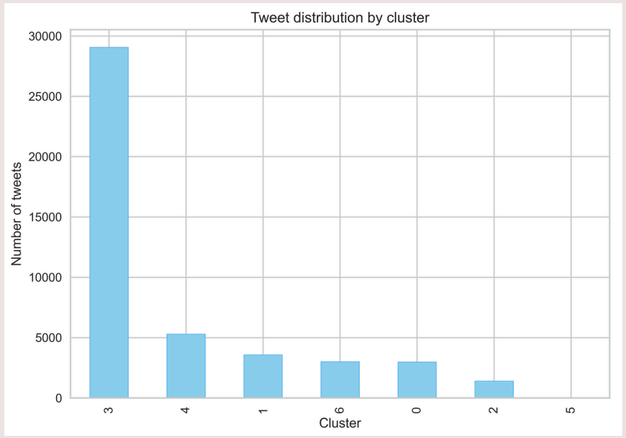
CLUSTERING

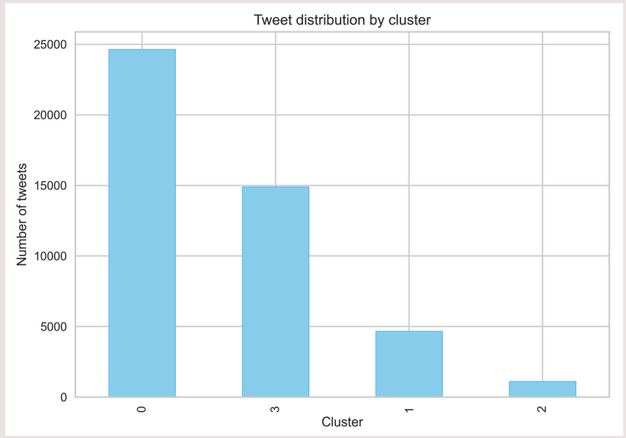
BoW

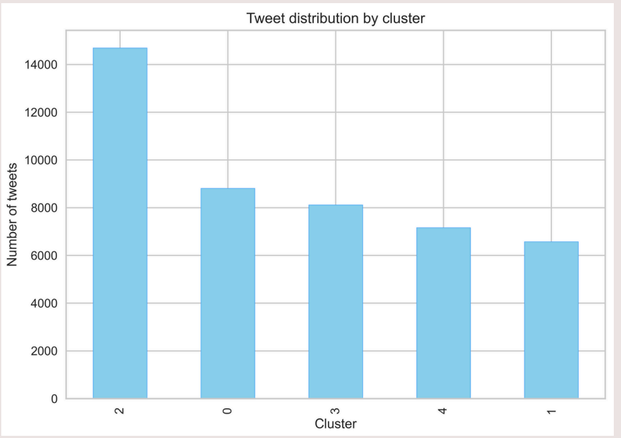
GloVe

sBERT

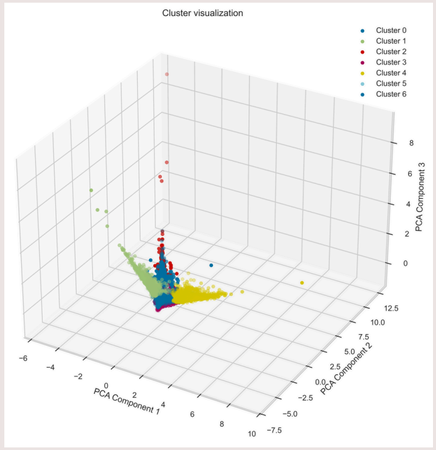
tweets distribustion

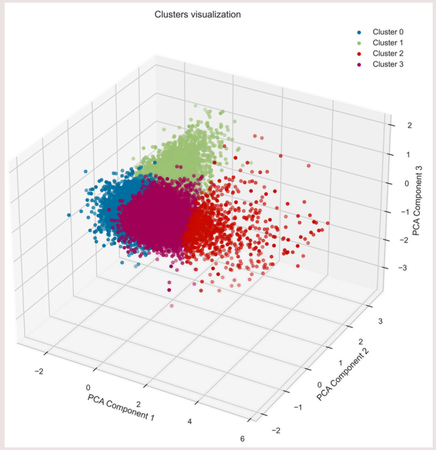


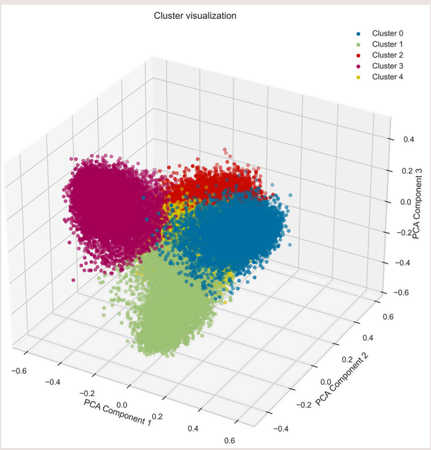




cluster visualization







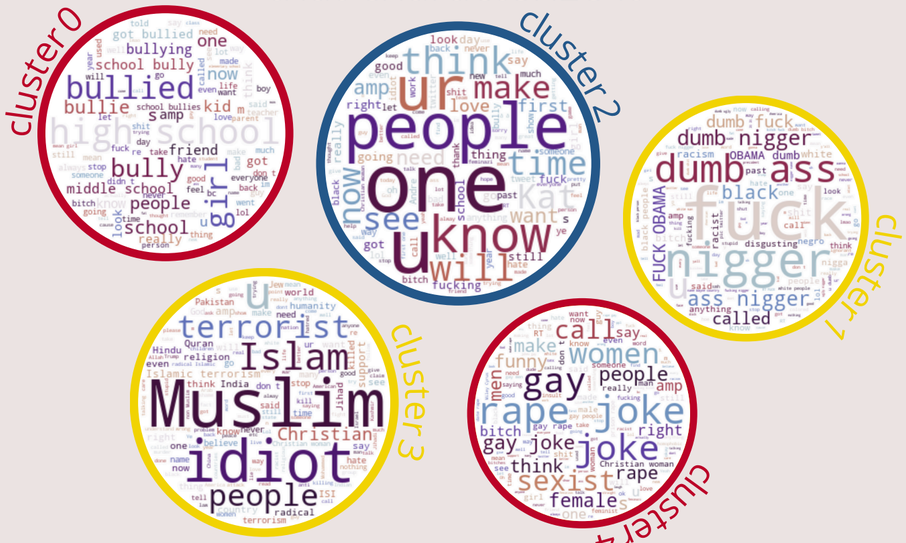
TOPIC MODELING

Our hypothesis: *sBERT helps to identify proper cluster membership with respect to ground truth*

How did we test this?

Compute the Silhouette score and number of cluster excluding one class at time

Analyze the word cloud for each specific cluster highliting the words that characterize it



CONCLUSIONS

Classification results tell us a story of native “censorship” within all models, a lazy solution even sBERT employs: this is what causes the cat and mouse game present nowadays on social media with controversial words

Clustering shows the importance of embeddings: only with contextual ones emergent properties such as the innate classification we apply on hate speech seems to arise

Topic modeling techniques confirm that indeed clusters reflect our original classes and uncovers the path to true moderation on social media, not through censorship but through better model able to understand context

THANK YOU

FOR YOUR ATTENTION