FRI Natural language processing course 2021
Project report

University of Ljubljana
Faculty of Computer and Information Science

Offensive language exploratory analysis

Benjamin Džubur, Žan Magerl and Jurij Nastran

Abstract

TBA

Keywords
Offensive language, clustering, visualizations, embeddings

Introduction

The use of offensive language and various offensive terms are often labeled with one or multiple of different categories. These categories are e.g. sexism, racism, hate speech, misogony, derailing, sexual harrasment, religion based, islamophobia, terrorism, etc. Notice how some of these categories are subcategories of others, some may be very closesly related or perhaps addressing the same issue. When data is labelled with some of these categories, it is usually done subjectively to a degree, and not by professional linguists. Therefore, the idea of this project is to use NLP and ML methods to analyze, explain and compare different categorizations of data, see what they consist of and how they relate to each other. We will concentrate on three datasets of tweets, two of which use multi-class categorization: CONAN [1] which includes 405 usable tweets and MMHS150K [2], containing around 150000 tweets. Lastly, the dataset prepared by Jha et al. [3] categorizes tweets into benevolent and hostile sexism.

Existing solutions

Many similar approaches have been performed on this and similar topics. Some notable are:

- Jaki & De Smedt [4] used skip-grams and k-means clustering on the top 250 most biased words of German hate speech tweets to identify clusters of words that are related.
- Rodriguez et al. [5] used TF-IDF representations of Facebook posts and related comments, which they then used with k-means clustering and extracted the most related words of the clusters and analyzed the topics.
- Reimers et al. [6] used agglomerative hierarchical clustering, k-means and DBSCAN clustering algorithms to cluster most similiar arguments. They studied how they can improve sentence-level argument similarity and clustering by using contextualized word embeddings.
- Sia et al. [7] compared two centroid based clustering algorithms on different word embeddings (Word2vec, ELMo, BERT, Fasttext etc.) in task of extracting top 10 words for different topics from news datasets.
- Malmasi et al. [8] used skip word bigrams and n-grams for word representation and then used Brown clustering in order to discriminate between hate speech and offensive speech.
- Ayo et al. [9] implemented a naïve Bayes model to improve feature representation and proceeded to use a modified Jaccard similarity measure for real-time probabilistic clustering of tweets into hate speech topic clusters.

Initial ideas

In order to use ML approaches such as clustering, we must represent words and documents numerically. We use different types of embeddings for this. A simple sparse embedding method is generating a TF-IDF matrix, using which we get simple representations for both words and documents. Another option is using neural, pre-trained word embeddings such as Word2vec to represent words, and document embeddings may be calculated based on averages of the containing words. For contextual embeddings such as BERT and ELMo, we can pass sentences of tweets as an input.

In order to group the most related words, we choose to use k-means clustering, agglomerative hierarchical clustering and DBSCAN because of their notably differing approaches. We analyse the results of the clustering methods on different kinds of embeddings and provide insights into how the choice of embedding effects the resulting clusters.

For dimensionality reduction, we will resort to t-SNE and

PCA, as they have proven to be the methods of choice for such data.

Concrete pipeline ideas

One basic approach which we will start with is the following. After preprocessing each tweet, we build TF-IDF embeddings, cluster the documents' embeddings using K-means with k equal to the number of target categories. Using different metrics, e.g. the Adjusted Rand Index we can compare how well the generated clusters align with the ground truth labels (given categories). Based on the clusters, we can also analyze the keywords they contain. Additionally, after dimensionality reduction, we can calculate distances and measure homogeneity of clusters when labelled with the given categories, to see which categories are related. Using t-SNE and PCA we project the data into 2-D and visualize the clustering results. Finally, we can merge the tweets of different datasets into one big dataset, and analyze how the different categorizations of each dataset might overlap (e.g. using Jaccardian index and similar scores) after applying the above algorithm.

The second approach is to group all tweets of the same category into a single document and perform keyword extraction on these documents. We can then use neural embeddings such as Word2vec on keyword of each category and use similarity measures to compare the different clusters and measure the similarity/distance between them.

A third approach would be to again use pretrained neural embeddings (Word2vec) to explore which words are in the neighborhood of the category terms, and see how well these words are represented in tweets of corresponding categories. Based on simple frequency measures of the filtered keywords in tweets, we can determine the uniqueness and sensibility of different categories. For each keyword we can also search the corpus and determine the most sensible/frequent context/category it is used in.

Finally, we can try generating BERT or ELMo embeddings on sentences or whole tweets and seeing if the added context information gives a more homogenous clustering result into the given categories than the first few approaches. Specifically, for the third dataset, where understanding the context of the use of sexist terms is crucial, we expect better results with this method.

Data gathering & preprocessing

As we have already mentioned, we will focus on three datasets. CONAN dataset contains islamophobic tweets, whose content is divided in many topics: culture, crime, rape, terrorism, sexism etc. After preprocessing the data, which included removing duplicates and tweets in Italian language, we've ended with 405 labeled tweets. MMHS150K dataset has no duplicated tweets, therefore there are almost 150 thousand tweets. Each tweet is labeled is labeled with up to 3 categories. Available categories are: racism, sexism, homophobia, religion and other. In the third dataset we have first removed duplicates and filter out the tweets that have been already deleted and we ended up with around 3350 tweets. Tweets are labeled with one of the two categories: hostile sexism or benevolent sexism.

In every tweet text we've removed punctuations and splitted sentences in the sequences of words. After that the next step was to remove common english stopwords. With the help of regular expression we've also removed tags of other users (denoted with @) and links to other pages. At the end we've have performed stemming on the remaining words.

References

[1] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy, July 2019. Association for Computational Linguistics.

[2] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications, 2019.

[3] Akshita Jha and Radhika Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, 2017.

[4] Sylvia Jaki and Tom De Smedt. Right-wing german hate speech on twitter: Analysis and automatic detection, 2019.

[5] A. Rodríguez, C. Argueta, and Y. Chen. Automatic detection of hate speech on facebook using sentiment and emotion analysis. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 169–174, 2019.

[6] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. Classification and clustering of arguments with contextualized word embeddings, 2019.

[7] Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too!, 2020.

[8] Shervin Malmasi and Marcos Zampieri. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202, 2018.

[9] Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, Idowu Ademola Osinuga, and Adebayo Abayomi-Alli. A probabilistic clustering model for hate speech classification in twitter. *Expert Systems with Applications*, page 114762, 2021.