Cyberbullying Detection in Social Media

Niv Shani (ID. 311361661)

Submitted as final project report for the NLP course,
Reichman University, 2023

Contents

1 Introduction

The rapid proliferation of social media platforms has brought both positive interactions and significant challenges. One of the most pressing issues is the rise of cyberbullying: a form of online harassment that involves the use of digital communication tools to intimidate, threaten, or demean individuals. The motivation for research in cyberbullying detection stems from the profound impact it has on individuals' mental well-being, self-esteem, and overall quality of life.

Detecting and addressing cyberbullying in social media requires a multidisciplinary approach that combines expertise from fields such as natural language processing, machine learning, psychology, and sociology. Researchers are developing advanced algorithms that analyze text data to identify offensive language patterns, hate speech, and harmful content. The challenge lies in distinguishing between genuine communication and false positives while considering linguistic nuances, cultural context, and evolving slang.

Overall, the field of cyberbullying is a very relevant and active field of research, that if properly mitigated can improve both our experience and use of social media, while increasing users' authenticity and reducing the overall hate and controversy across society.

2 Solution

2.1 Datasets

- Offensive Language Identification Dataset (OLID) [1]

  Each tweet is annotated for different offensive categories, in three levels:

  - Level A: offensive / not offensive (binary classification)
  - Level B: offense type (if offensive) - targeted / not targeted.
  - Level C: the target type (if targeted) - individual / group / other (an organization, an event, an issue, etc.)

- Donald Trump's tweets collection before and during his US presidency [2]

2.2 General approach

The general idea of this project is to train an offensive tweets classification, for the purpose of identifying trending targets that are potentially bullied in social media. Given a dataset of tweets, we can classify them and extract the *targeted* offensive tweets (as opposed to general rude language or content), then analyze the frequent subjects (considering specific POS tags, single words or bigrams, etc.) This will highlight the trending subjects of cyberbullying for that given dataset, an ability that can be developed and leveraged for content moderation, protection, and overall mitigation of the offensive agendas in social media.

2.3 Design

2.3.1 Tweets classification - `TweetsBullyingClassifier`

Originally, I planned to train an HMM-based classification model over the labeled *OLID* dataset, and extract entities or POS tags from the tweets. I struggled to gather labeled training data for POS tagging of this specific task,

so I've decided to shift the classification task to using a *Logistic Regression*[3] based classification model.

Features extraction

As a pre-processing step, the tweets have been cleaned up (striped user tags, hashtags, URLs and stop-words). Regardless, the following features were extracted for the classification task:

`tweet_length`: number of characters in the tweet.

`words_count`: number of words in the tweet.

`hashtags_count`: number of distinct hashtags in the tweet.

`emojis_count`: number of distinct emojis in the tweet.

`tags_count`: number of distinct handles in the tweet (e.g., *@USER*)

`exclamation_mark_count`: number of exclamation marks in the tweet.

`all_caps_count`: number of capital letters in the tweet.

`digits_count`: number of digits in the tweet.

`has_url`: indicate whether the tweet contains an external link or not.

`sentiment`: tweet sentiment analysis.

`bow_words_prediction`: tweet class prediction based on words frequency.

`bow_hashtags_prediction`: tweet class prediction based on hashtags frequency.

2.3.2 Sentiment analysis - `SentimentAnalyzer`

I've used a shelf pipeline[4] for tweets sentiment analysis, to be used as a classification feature for the `TweetsBullyingClassifier`.

The classifier scores the tweet sentiments and produces one of three labels (the highest score): **POS**: positive sentiment, **NEU**: neutral sentiment and **NEG**: negative sentiment. The tweets sentiment analysis was part of the data pre-processing and took approximately 3-3.5 hours. An output mapping `tweet_uid` → `sentiment` was stored in the attached `sentiments.csv` file which is loaded during the classifier initialization.

2.3.3 Frequent terms analysis - `WordFrequencyClassifier`

I've trained a *TF-IDF*[5] model (Term Frequency–Inverse Document Frequency) on the training data for analyzing the frequent terms (words and hashtags) in the tweets. Common English stop words[6] were skipped, leaving mostly nouns and pronouns as frequent subjects for hashtags and content.

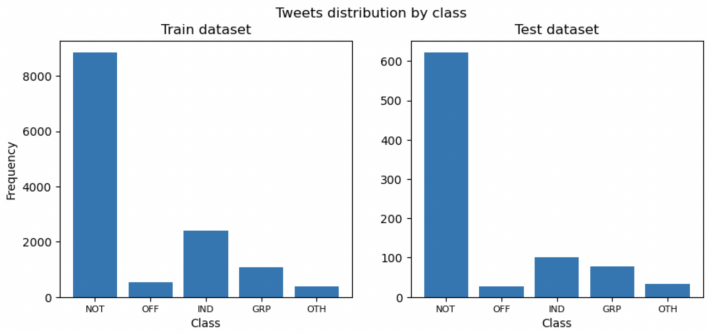
The model then classifies a given tweet to one of the five original labels. These two predictions (frequent words and hashtags classification) is then used as two separate classification features for the `TweetsBullyingClassifier`.

3 Experimental results

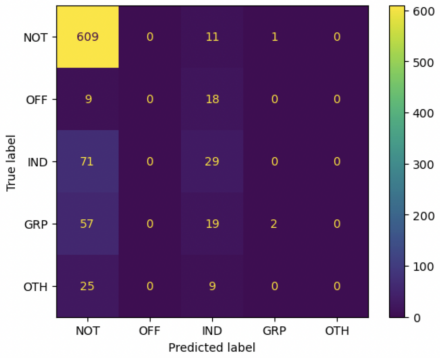
The model outputs numeric labels (1-5):
1 (*NOT*): Non-offensive, **2** (*OFF*): Offensive but not targeted, **3** (*IND*): Offensive and targeted toward an individual, **4** (*GRP*): Offensive and targeted toward a group or organization, and **5** (*OTH*): Offensive and targeted toward other subject (idea, location, etc.)

3.1 Model performance

The non-offensive tweets (labeled **1**) has a much more prominent presence in the training data, which can explain the model's overall accuracy scores. These results can be significantly improved by training the model on a larger set of offensive and targeted tweets.



	precision	recall	f1-score	support
1	0.79	0.98	0.88	621
2	0.00	0.00	0.00	27
3	0.34	0.29	0.31	100
4	0.67	0.03	0.05	78
5	0.00	0.00	0.00	34
accuracy			0.74	860
macro avg	0.36	0.26	0.25	860
weighted avg	0.67	0.74	0.67	860



4 Discussion

As mentioned before, the original goal of the project was **trends extraction** - the ability to extract and identify offensive trends from a given tweets dataset.

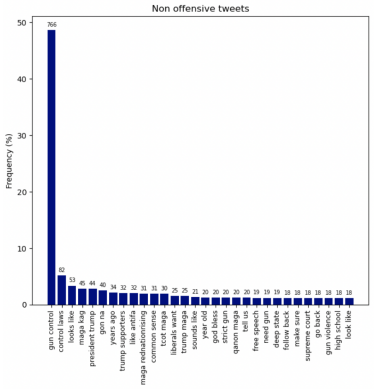
Here are two different applications of the model I've described above that achieves this goal, to some extent. Clearly, this model can be further fine-tuned and vastly trained for better results:

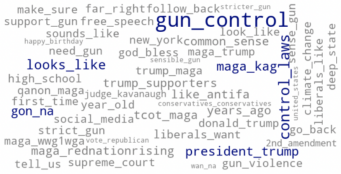
- Explore the tweet's author, his communities or interests, for extracting more features, and eventually increase the model accuracy. Gathering a more balanced dataset (with more offensive and targeted samples) would probably also help to increase the model accuracy over these classes.
- Train for a specific task on specific datasets, jargons or slang.
- Explore longer n-grams or different named entities.

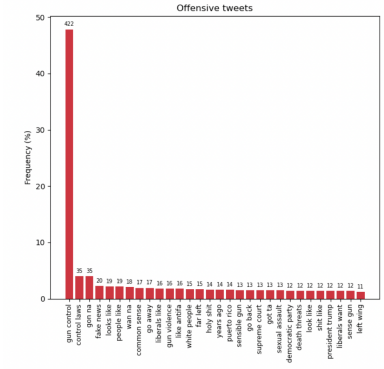
4.1 Offensive vs. non-offensive tweets (training and test datasets)

For this demonstration, we only look at the tweets with the binary question of offensive or not, regardless of any potential targeting classification. Analyzing the frequent *bigrams* of the dataset, we can clearly see some trends:

- The terms `gun control` and `control laws` are clearly widely discussed topics which are common to both offensive and non-offensive tweets.
- The term `fake news` is unique to offensive tweets - this can be further investigated to analyze the typical user persona, relevant hashtags or general discussions.





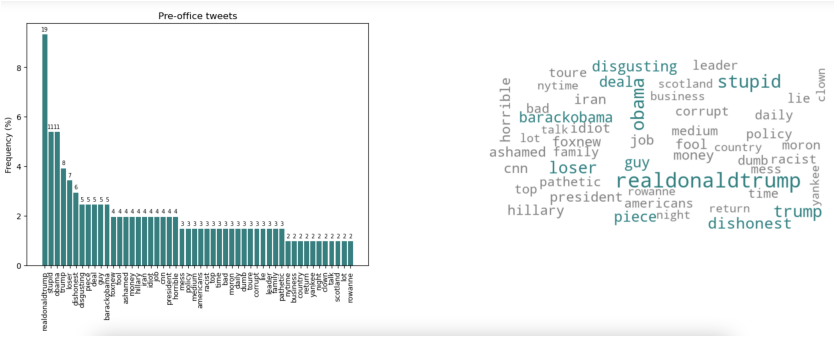


Filtering the dataset to look specifically at the tweets labeled as *targeted* (either towards individuals, groups or other subjects) - we can see that a new bigram is frequent - `white people`.

4.2 Donald Trump's pre-office vs. in-office tweets

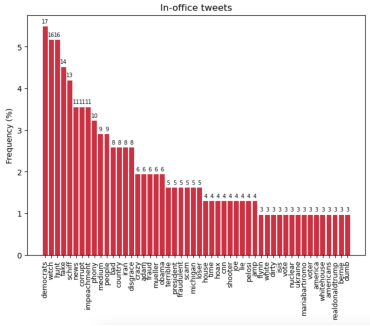
Using a public dataset[2] of Trump's Twitter user (`@realDonaldTrump`) until it was banned from Twitter, we're looking at tweets from his time before the US presidency, and his tweets while he was the US president. Applying the model, I've classified 5000 pre-office and 5000 in-office tweets, then filtered to pick the targeted ones.

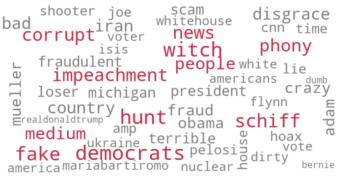
While not as significant at this dataset size, we can clearly see the change of trending subjects Trump is referring to during his election campaign (pre-office) and after he was elected. Pre-office, the most frequent trend in his offensive tweets was to slander his political rivals or his opposers - we can see terms like `obama`, `hillary`, `loser`, `stupid` and `clown`.



On the other hand, applying the same point of view on Trump's in-office tweets,

we can see his tone was directed much more towards affecting public opinion, accusing and dealing with criticism - we can see terms like `witch hunt`, `fake news`, `hoax`, `democrats` and `impeachment`.





5 Code

Full GitHub repository (notebook, datasets and relevant files included):
https://github.com/Nivsha08/nlp-2023-final-project

6 References

1. Offensive Language Identification Dataset (OLID)
2. Complete Trump Tweets
3. sklearn.linear_model.LogisticRegressionCV
4. finiteautomata/bertweet-base-sentiment-analysis
5. sklearn.feature_extraction.text.TfidfTransformer
6. nltk.corpus.stop_words