Group Name: individually
Name: Kristina Kaliagina
Email: kalyagina.kristina@gmail.com
Country: Russia
College/Company: Graduated from the University “Higher School of Economics”
Specialization: NLP

Deliverables

Problem description:

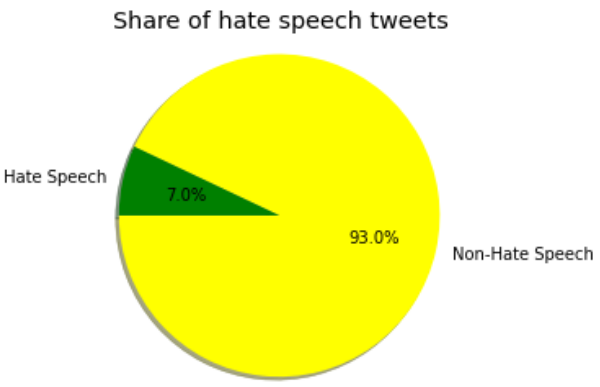
The task of the project is to classify tweets from Twitter, namely, it is necessary to create a model that will help determine whether a particular tweet belongs to such a type of speech as hate speech. Hate Speech, in simple terms, is offensive language directed at individuals or groups based on their affiliation, interests, and characteristics, such as their religion, nationality, race, color, origin, gender, or other identity factor.

The task is quite difficult because of the inherent complexity of natural language constructs - different forms of hatred, different types of goals, different ways of representing the same meaning.

Data understanding:

To solve the problem of determining hate speech, data was taken in the form of tweets from the well-known Twitter platform. The data consists of a training and test set. In total, these are 49159 observations, 31961 and 17197, respectively.

The training sample, in turn, has a dependent variable called "label", which takes on the value 1 if the tweet refers to hate speech and 0 if it does not. The data is very unbalanced, as 7% of the training sample belongs to the hate speech.



In addition, tweets have some noise. The tweets contain hashtags, other users' names, greek characters, interjections, slang, and so on.

Type of data for analysis:

The project data consists of 2 csv files: training data and test data.
There are two types of data in data files: int (features “id” and “label) and object (feature “tweet”).

Data problems:

- Unbalanced classes
- Duplicate retweets
- Passes, extra spaces
- Lots of punctuation
- @usernames and #hashtags
- Slang
- Interjections
- Greek symbols

Approaches to solve the problems in the data:

1. Tokenization - separating tweets by sentences

2. Lematization/ Stemming - bringing words from tweets to their root form to reduce variations of the same word

3. Removing stop words - these are irrelevant words that will not help identify the text of the tweets as a hate speech or non-hat speech

4. Regular expressions - for example, to remove punctuation marks as they don't add value or meaning to the NLP model, remove hashtags, usernames and so on

5. Correction of misspelled words - replace slang with ordinary words that are accessible to everyone

6. Bag of words - convert text into sets of numbers (vectors), as machine learning algorithms require this

7. N-grams - create a dictionary from a sequence of words to have more ideas about the data

8. TFIDF - estimating the importance of a word in tweets