Toxic Comment Classification

Proposal

1 Domain Background

The use of different online platforms (social media, messaging platforms, gaming platforms and mobile phones) has become a part of our everyday life. Negative consequences for mental health could include exposures to online forms of aggression. Cyberbullying is an emerging issue in times of digital technologies. Since adolescent public makes an extensive use of online platforms, they are affected by cyber harassment the most.

Cyberbullying can take on many forms, including personal attacks, harassment or discriminatory behavior, spreading defamatory information, misrepresenting oneself online, spreading private information, social exclusion and cyberstalking [1]. Pew research center [2] and Cyberbulling research center [3] conducted research which involved US teens and now provide more data and visualisations on subject matter.

The comment sections of social media and news outlets have become the new playground for online bullying. A research shows that 1 out of 5 online harassment takes place in comment space [4]. Keyboard courage, driven by anonymity and the absence of real world response, may ruin the experience of a lot of users. As a result, many news organizations are even choosing to eliminate comments altogether to avoid this problem.

Users' feedback becomes quickly unmanageable if an online community grows fast, manual moderation of comments seems no longer an option, automatic tools should be used to detect and block toxic comments. Recent developments in natural language processing, as well as in machine and deep learning, are able to provide such solutions.

2 Problem Statement

The goal of this project is to build a classification model that allows to detect different types of toxicity (obscenity, threats, insults, and identity-based hate).

The initial project and data come from the Conversation AI team, a research initiative founded by Jigsaw and Google (both a part of Alphabet). They are working on tools to help improve online conversations. One area of focus is the study of negative online behaviors, like toxic comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion) [5].

Automatic toxic comment detection is an important issue in modern society, since people use social media worldwide. Recently, Jigsaw released data for Multilingual Toxic Comment Classification [6], similar competitions are available on Kaggle for NLP enthusiasts. Scholars study this topic as well, developing models and tools to make social media a safe place for non-violent and respectful communication [7] [8] [9] [10].

3 Datasets and Inputs

The data[1] is available on Kaggle competition page:
https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data.

[1]The dataset is under CC0, with the underlying comment text being governed by Wikipedia's CC-SA-3.0.

3 files are in our disposal: `train.csv`, `test.csv`, `test_labels.csv` (cf. attached images showing first 10 lines in each file):

1. `train.csv` - training data, contains comments with their binary labels;
2. `test.csv` - the test set, containing only comments (Note: to deter hand labeling, the test set contains some comments which are not included in scoring);
3. `test_labels.csv` - labels for the test data (Note: value of -1 indicates it was not used for scoring).

These files provide a large number of Wikipedia comments which have been labeled by human raters for toxic behavior. There are 6 types of toxicity:

- `toxic`,
- `severe_toxic`,
- `obscene`,
- `threat`,
- `insult`,
- `identity_hate`.

Besides, there is a large number of comments, which are not tagged with any of toxic labels, i.e. neutral. We will deliberately introduce a `neutral` label for further data analysis, as well as for classification[2]. The overall data distribution across training set is displayed in Table 1.

neutral	**toxic**	**severe toxic**	**obcsene**	**threat**	**insult**	**identity hate**
143346	15294	8449	7877	1595	1405	478

Table 1: Distribution of data in training set

Table 1 shows that the data is highly unbalanced: 80% of all the comments in training set are neutral (i.e. do not have any associated toxicity labels), whereas all the 6 classes of toxicity are distributed among the rest of 20% of the data.

4 Solution Statement

To address the problem of automatic detection of toxic comments we will first build a simple and efficient Naive Bayes classifier, an algorithm which is commonly used in text classification. Classifiers built with machine learning tools, such as Naive Bayes, perform class labels assignment based on a set of features which can be represented numerically with statistical methods.

However, since we deal with text data, LSTMs seem to be a more appropriate choice for this type of input, pretrained embeddings are also commonly used in this case. We will explore them both as our main model. An LSTM (Long-Short-Term-Memory classifier) is a recurrent neural network which is able to learn order dependence in sequence problems; word embeddings are vector representations of words where words with similar meaning have similar vectors.

Given that the data is disproportional due to the high number of neutral comments, we will explore two possibilities:

[2]Including `neutral` label is not required by Kaggle competition; however, for the preset project we only take the relevant data from it; we will explore and model it regardless of official competition requirements.

1. Build multilabel classifiers:
    - with `neutral` label included - highly unbalanced data,
    - with `neutral` label excluded - more homogeneous classes;
2. Build binary classifiers:
    - to classify toxic vs. non-toxic comments - unbalanced data,
    - to classify toxic vs. non-toxic comments, with a SMOTE oversampling technique - to reduce disproportions in data.

We will furthermore contrast both models performances (Naive Bayes and LSTMs) for these four types of classification problems.

5 Benchmark Model

For a benchmark model we will construct a Multinomial Naive Bayes classifier (build with TfIdf vectorizer) for 7 output labels (`neutral`, `toxic`, `severe_toxic`, `obscene`, `threat`, `insult`, `identity_hate`).

The Naive Bayes algorithm is frequently used as a benchmark for different classification tasks.

6 Evaluation Metrics

We will use several metrics to assess the performances of all the models:

- Accuracy score (fraction of correct predictions),
- Precision (the ability of the classifier not to label as positive a sample that is negative) and recall (the ability of the classifier to find all the positive samples),
- ROC AUC score (fraction of true positives out of the positives, true positive rate, vs. the fraction of false positives out of the negatives, false positive rate),
- Confusion matrix: (the exact numbers on TPs, TNs, FPs, FNs).

We will study the results by analyzing all these metrics: we hypothesize that, since our data is unbalanced, the first three metrics might be very different; however, ROC AUC might be the most reliable metric given data percularities.

7 Project Design

The project will be designed as follows:

1. We will first load and unpack `.zip` files with all the necessary data (for training and test sets, as well as for embeddings - we will work with Glove);
2. We will then proceed with data diagnostics, such as checking for duplicates and/or missing values;
3. We will perform exploratory data analysis:
    - assessing training and test shape,
    - visualizing some lines from both sets to get an overview of the data,
    - perform calculations on data distribution: number of comments classified with each label (cf. Table 1); number of comments per number of labels;
4. The next step will be to preprocess the raw text data using NLTK tools:

- lowercasing and removing punctuation,
- splitting comments into words,
- removing stopwords;

5. Using the output of preprocessing step we will be able to visualize the most frequent words appearing under each label with word clouds;
6. The final step of data processing will be to transform textual information into numerical:
   - TfIdf vectorizer will be used to prepare data for Naive Bayes classifier,
   - TensorFlow tokenzer and padding sequences tool will be used to prepare data for LSTMs; at this stage we will also construct an embedding weights matrix using Glove pretrained embeddings;
7. We will proceed with modelling:
   - The first task will be to construct a benchmark model - multinomial Naive Bayes classifier for 7 output labels,
   - We then will use LSTMs for this task,
   - Next, we will construct both multilabel classifiers - removing `neutral` label from outputs,
   - We will continue with binary classification models - to classify toxic and non-toxic comments,
   - The final models will perform the same binary classification - applying SMOTE oversampling technique;
8. Finally, we will analyze and compare the results of these classification models using evaluation metrics;
9. We will conclude with final remarks and outline other possible solutions to address the toxic comment classification problem.

8 References

[1] Social media addiction linked to cyberbullying:
https://news.uga.edu/social-media-addiction-linked-to-cyberbullying/

[2] A Majority of Teens Have Experienced Some Form of Cyberbullying:
https://www.pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/#fnref-21353-1

[3] 2016 Cyberbulling data:
https://cyberbullying.org/2016-cyberbullying-data

[4] About 1 in 5 victims of online harassment say it happened in the comments section:
https://www.pewresearch.org/fact-tank/2014/11/20/about-1-in-5-victims-of-online-harassment-say-it-happened-in-the-comments-section/

[5] Toxic Comment Classification Challenge:
https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[6] Jigsaw Multilingual Toxic Comment Classification:
https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification

[7] Identification and Classification of Toxic Comments on Social Media using Machine Learning Techniques:
https://www.rsisinternational.org/journals/ijrias/DigitalLibrary/Vol.4&Issue11/142-147.pdf

[8] Learning from Bullying Traces in Social Media:
https://aclanthology.org/N12-1084.pdf

[9] Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit:
https://www.mdpi.com/2079-9292/10/11/1332/htm

[10] The Distorting Prism of Social Media. How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity:
https://cpb-us-e1.wpmucdn.com/sites.dartmouth.edu/dist/5/2293/files/2021/03/comment-toxicity.pdf