FRI Natural language processing course 2021
Project report

University of Ljubljana
Faculty of Computer and Information Science

Cross-lingual offensive language identification

Gojko Hajduković, Simon Dimc

Abstract

To be added.

Keywords
Text classification, Hate speech classification

Advisors: Slavko Žitnik

Introduction

Text classification is one of the most important and well researched tasks in Natural Language Processing. It represents a supervised learning approach whose goal is to classify an input text into the correct category, therefore each input example gets assigned a label and as such is used in wide range of applications. Applications that exploit text classification the most are document classification, web search, information retrieval, sentiment analysis and spam detection. Text classification breaks down to different levels of scope:

1. **`Document level`** - In the document classification algorithm, classifies document as a whole into the categories
2. **`Paragraph level`** - Classifies a portion of a document, a single paragraph into the required categories.
3. **`Sentence level`** - Obtains the categories of a portion of a paragraph, a single sentence.

In this paper we focus on online hate speech classification problem. Online hate speech classification has become a well researched part of NLP and an urgent need since human supervision is unable to handle an enlarging volume of online content. We further divide classification problem based on the number of target variables to be classified. At first we classify input text at a paragraph level into binary categories, while in the later phase we classify input text on multi-class labels. In this paper we will explore two types of models for binary classification problems. First, traditional machine learning models such as `Logistic regression`, `SVM`, `Random forest` and an `Ensemble voting model` are to be employed into classifying online hate speech. The later approach will rely on `BERT` an transformer based machine learning technique. In case of successfully validating our initial hypothesis of correctly classifying hate speech we will introduce a Slovene labeled hate speech data set consisting of 200 comments. Additionally, we will transfer the knowledge from pre-trained multi-lingual models using one-shot transfer and evaluate introduced Slovene data set.

Related work

With the enlarged amount of online content, the need and the and focus of research has shifted towards online text classification, with a particular focus on content moderation. Within the early stages of text moderation number of methods utilizing shallow approaches have been proposed. Dao et al. [1] introduced methods wich use feature vectors constructed with BoW language model. The main shortcomings of the introduced methods is that it suffers from data sparsity and that it typically leads to many false positives. Other approaches for feature spaces comprising of TF-IDF and character n-grams have been introduced. Waseem et al. [2] explore the traditional, feature-engineered machine learning models such as Logistic Regression and SVM for automatic content moderation.

In order to tackle the problem of data sparsity, number of approaches such as Word2Vec[3] and GloVe [4] have been introduced for better representation of words and sentences. These methods outshine sparse feature vectors since similar words are located closer together in the latent space and dense word representation produce better encoding of such features. Recent studies show that Neural Network based approaches are superior compared to the traditional machine learning methods. Advantage of neural network approaches lays in it's capability to model larger sequences of text. Sutskever et al. [5] show that gated RNNs are capable of capturing long term dependencies in text leading to a better results in content classification.

The main shortcoming of NN based approaches lays in it's inability to catch global context of the words. A breakthrough in the field of NLP happened with introduction of BERT. Bidirectional Encoder Representations from Transformers - BERT represents a SOTA approach in NLP. Compared to the other dense word representations BERT takes into account a larger left and right semantic space thus better modelling words with respect to the context around it. There are a number of pre-trained BERT models which can be fine tuned on various downstream NLP tasks. Furthermore, there are a number of available multilingual pre-trained BERT models, thus allowing a transfer of model knowledge to the less resourced languages,

Methods

In this section we introduce data sets used for training and testing, preprocessing steps as well as all the performed experiments.

Data

As aforementioned, we have subdivided speech classification task into two categories, a binary and a multi class label text classification. For both task categories we have used english and slovene data sets. For binary task we have used four publicly available english and two slovene data sets that have been further processed and adopted in order to be combined into one uniform binary class data set. The dataset consists of hate-speech labelled with **1** while non-hate speech is annotated with **0**. While, for mutliclass task we have used two english data sets out of which one is publicly available as well as two slovene datasets which are then processed and adopted into one uniform multiclass dataset.

Binary english datasets

All the below listed four datasets have been processed and adopted to form one concatenated dataset. The concatenated dataset consisted of 67575 examples out of which 4541 were duplicates which have been removed, thus the final dataset consists of 63034 examples. The dataset is highly imbalanced consisting of 42697 non-hateful comments and 20337 hateful ones.

Dataset 1

First dataset represents a 1528 Fox news comments [6] which are manually classified into hate and non-hate categories. publicly available at url. The original data set is a json file where each comment has an additional metadata and previous contextual comments. All the metadata have been removed, preserving only the text and the binary label. Table 1 shows that original dataset is highly imbalanced consisting of 435 hateful and 1093 non-hateful comments.

Dataset 2

Second dataset represents a 26581 comments [7] from Reddit posts which are manually classified into hate and non-hate categories, publicly available at url.The original data set is a csv file where each comment has an additional metadata and previous contextual comments. All the metadata have been removed, preserving only the text and the binary label. Table 1 shows that original dataset is highly imbalanced consisting of 5225 hateful and 21356 non-hateful comments .

Dataset 3

Third dataset represents a 38018 comments [7] from Gab posts which are manually classified into hate and non-hate categories, publicly available at url.The original data set is a csv file where each comment has an additional metadata and previous contextual comments. All the metadata have been removed, preserving only the text and the binary label. Dataset consists of 14139 hateful comments and 23888 non-hateful comments.

Dataset 4

Fourth dataset represents a 24802 comments [8] from Twitter posts which are manually classified into hate, offensive and neither categories publicly available at url. Offensive comments have been filtered out from the original dataset resulting into a csv shaped dataset consisting of 1430 hateful comments and 4163 non-hateful ones.

Multiclass english datasets

The two below listed datasets consist of multi-topic harrasment comments. All the comments have been extracted, processed and concatenated with the non-hateful ones from aforementioned binary data in order to form unified multi-class data set. The dataset consisted of 59261 comments out of which 4039 were duplicated which have been removed, thus the final dataset consists of 55222 examples distributed across 6 label categories. The table 2 shows number of examples across each category as well as label codes.

Dataset 1

First multiclass dataset consists of 10236 comments [9] from Twitter posts which are manually classified into 7 categories showed in table 4. The dataset is publicly available at url. The dataset is processed and labels are combined according to table 4 in order to be concatenated into final multiclass data set.

Dataset 2

Second multiclass dataset consists of 2956 comments [10] from Twitter posts which are manually classified into 5 categories showed in table 4. The dataset is processed and labels are combined according to table 4 in order to be concatenated into final multiclass data set.

Preprocessing

Preprocessing is applied to all the instances of each dataset that the classifier could predict class for. Preprocessing is conducted through eight steps. Since our datasets are collected mainly from the social media or news media comments there are high number of occurences of emoticons. With step

1 all the emoticons are removed. Same reasoning is behind step 2 and 3 where all twitter urls, hashtags and @name mentions are removed. Steps 4 and 5 remove all non-alphabetical characters and lower down all words. Step 6 tokenizes each input comment with nltk function word_tokenize which firsts tokenizes text into sentences than it further tokeniizes it into tokens. Lemmatization is performed in step 7 in order to get word in a canonical form and get higher frequency of similar words. Nltk WordNetLemmatizer is used for converting a word into lemmas. Additionaly, step 8 combines english stop words removal as well as filtering out words of length 1.

Features

For the feature extraction in our baseline model we have used sparse-word embeddings. First, we have used simple frequency count to create feature sets. In the later approach we employed tf-idf technique. With both approaches we have used unigrams and bigrams while using Sklearn implementations of **CountVectorizer** and **TfidfVectorizer** respectively.

Models

Table 1 and 2 show that data sets are highly imbalanced which can lead to classification algorithms having low accuracy towards the minority class
es. In order to tackle the problem of imbalanced datasets we have tested under-sampling and over-sampling on training data as well as adjusting class weights.

Traditional ML models

For the baseline models we have selected SKlearn implementation of Logistic regression, Support-Vector Machine and and Ensemble voting classifier - XGBoost.
Logistic regression - We have used sklearn implementation of Logistic Regression classifier. For the model parameter tuning the GridSearchCV is used to choose the best model for predicting classes of each dataset as well as predicting classes for a dataset that is a concatenation of all datasets. Table [] shows the best choice of parameters used for predicting classes.
Support-Vector Machine - We have used sklearn implementation of Support-Vector Machine classifier. For the model parameter tuning the GridSearchCV is used to choose the best model for predicting classes of each dataset as well as predicting classes for a dataset that is a concatenation of all datasets. Table [] shows the best choice of parameters used for predicting classes.

Ensemble models

XGBoost - We have used XGBoost ensemble model for classifying comments data. The boosting algorithm is used with a learning_rate = 0.05, initial n_estimators=300.

Transformer ML models

BERT - In order to utilize BERT for text classification we have used BertForSequenceClassification model with classification head from https://huggingface.co/transformers/ library. In the experiments classification head is based on either mBert or CroSloEngualBert. Both models are pre-trained on a large corpora of multilingual data and are suitable for cross-language knowledge transfer. The model fine-tuning is powered by https://pytorch.org/ framework.

Table 1. Binary-class dataset label distribution.

Non-hate	Hate
47055	21502

Table 2. Multi-class dataset label distribution.

Non-hate	Race	Sexual	Intelligence	Appear	Oth.
42831	6414	3769	1197	653	358

Experiments and Results

Traditional ML models

In the first pass all the models are trained and evaluated on each individual english data set for both, binary and multiclass task. The process of training and testing is conducted via stratified K-fold validation with $K = 5$ on both feature sets composed from BoW or tfidf. In the each pass, the data is split on training and testing data with $80-20\%$ ratio. In the second pass the same procedure is repeated on afore described combined dataset for both binary and multiclass tasks.

Transformer ML models

References

[1] Y. Wang I. Kwok. Locate the hate: Detecting tweets against black. *AAAI*, 2013.

[2] Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. *NAACL-HLT, pages 88–93*, 2016.

[3] K. Chen G. S. Corrado J. Dean T. Mikolov, I. Sutskever. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems, pages 3111-3119*, 2013.

[4] C. D. Manning J. Pennington, R. Socher. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing, pages 1532-1543*, 2014.

Table 3. Model classification weighted f1-score on binary class labels

Imbalanced					
DS 1	DS 2	DS 3	DS 4	DS 1-4	Model
0.739	0.845	0.792	0.934	0.87	LR
0.774	0.865	0.823	0.912	0.897	SVM

Table 4. Multi-class dataset label mappings

DS1	Final DS	DS2	Final DS
Archaic	Racism	Racial	Racism
Ethnicity	Racism	Appearance	Appearance
Nationality	Racism	Intelligence	Intelligence
Class	Intelligence	Political	Other
Disability	Intelligence	Sexual	Sexual
Gender	Sexual		
Sex	Sexual		
Religion	Other		

[5] Q. V. Le I. Sutskever, O. Vinyals. Sequence to sequence learning with neural networks. *Advances in neural information processing systems, pp. 3104–3112*, 2014.

[6] L. Gao and R. Huang. Detecting online hate speech using context aware models. *ArXiv*, 2018.

[7] Bethke A. Belding E. Qian, J. and W. Yang Wang. A benchmark dataset for learning to intervene in online hate speech. *ArXiv*, 2019.

[8] Automated hate speech detection and the problem of offensive language, year = 2017, journal = ArXiv, author = Davidson, T., Warmsley, D., Macy, M. and Weber, I,.

[9] Nilizadeh S. Nguyen D. Vigna G. ElSherief, M. and E Belding. Peer to peer hate: Hate speech instigators and their targets. *Proceedings of the Twelfth International AAAI Conference on Web and Social Media, pp.52-61*, 2018.

[10] Shekarpour S. Balasuriya L. Thirunarayan K. Shalin V. Rezvan, M. and A. Sheth. A quality type-aware annotated corpus and lexicon for harassment research. *ArXiv*, 2018.

Table 5. Model classification weighted f1-score on multi class labels

Imbalanced			
DS 1	DS 2	Combined DS	Model
0.931	0.897	0.929	LR
0.942	0.901	0.943	SVM