Course Project

Information Retrieval and web Search

By

INFOMEN

Abhishek Singh : 2018CH10188

Saurav Mittal : 2018CH10243

Ishaan Garg : 2018TT10905

Project Checkpoint 2 : *Detailed project description, approach taken, resources needed, datasets and metrics etc*

Project Title :

Hate Speech and Offensive Content Identification in Indo-European Languages

Problem Overview :

A significant threat to societies is the vast fraction of hate speech and other inflammatory and inappropriate content online. Objective discussions are compromised by offensive language such as disrespectful, hurtful, degrading, or obscene material directed from one individual to another person and available to others. Such kinds of language can be seen more and more on the web and can lead to radicalized conversations. The development of public opinion requires logical, critical discourse. Simultaneously, without enforcing rigid censorship regimes, open societies need to find an appropriate way to respond to such content. As a result, several social media website sites track user messages. This leads to a pressing demand for methods to recognize suspicious posts automatically. To deter harmful activity in social media, online communities, social media platforms, and technology businesses have been investing extensively in technology and processes to detect offensive language.

We take this problem on hand and try to provide a viable solution, taking social media content in different Indo-European languages into consideration.

Project Description :

The identification of Hate Speech in Social Media has received much attention in research recently. For languages other than English, there is a particular demand for research. In Hindi, German, and English, Hate Speech Identification resources created by HOSAC have been used. Three Twitter and Facebook datasets have been created and made available. We intend to stimulate Hate Speech classification research and development for various languages. The datasets enable supervised machine learning systems to be developed and tested. In 3 sub-tasks, binary classification and more fine-

grained sub-classes are offered. Models based on deep learning methods have proved to be sufficient for the classification task.

There has been significant work in several languages in particular for English. However, there is a lack of research on this recent and relevant topic for most other languages. The objective here is to stimulate research for these languages and to find out the quality of hate speech detection technology in other languages.

The HASOC dataset provides several thousand labelled social media posts for each language. The entire dataset was annotated and checked by the organizers of the track. The annotation architecture is designed to create data for 3 different sub tasks.

1. SUB-TASK A: classification of Hate Speech (HOF) and non-offensive content.

2. SUB-TASK B: If the post is HOF, sub-task B is used to identify the type of hate.

3. SUB-TASK C: it decides the target of the post.

Task Description

Sub-task A : Sub-task A focuses on Hate speech and Offensive language identification and is offered for English, German, Hindi. Sub-task A is coarse-grained binary classification in which participating system are required to classify tweets into two class, namely: Hate and Offensive (HOF) and Non-Hate and offensive.

1. (NOT) Non Hate-Offensive - This post does not contain any Hate speech, offensive content.

2. (HOF) Hate and Offensive - This post contains Hate, offensive, and profane content

Sub-task B : Sub-task B represents a fine-grained classification. Hate-speech and offensive posts from the sub-task A are further classified into three categories.

1. (HATE) Hate speech: Posts contain Hate speech content.

2. (OFFN) Offensive: Posts contain offensive content.

3. (PRFN) Profane: These posts contain profane words

HATE SPEECH : Describing negative attributes or deficiencies to groups of individuals because they are members of a group (e.g. all poor people are stupid). Hateful comment toward groups because of race, political opinion, sexual orientation, gender, social status, health condition or similar.

OFFENSIVE : Posts which are degrading, dehumanizing, insulting an individual, threatening with violent acts are categorized into this category.

PROFANITY : Unacceptable language in the absence of insults and abuse. This typically concerns the usage of swearwords (Scheiße, Fuck etc.) and cursing (Hell! Verdammt! etc.). Such posts are categorized into this category. As expected, most posts are in the category NOT, some are HATE and the other two categories are less frequent. Dubious cases which are difficult to decide even for humans, were left out.

Sub-task C (only for English and Hindi) : Sub-task C considers the type of offense. Only posts labeled as HOF in sub-task A are included in sub-task C. The two categories in sub-task C are the following:

1. Targeted Insult (TIN): Posts containing an insult/threat to an individual, group, or others.

2. Untargeted (UNT): Posts containing non targeted profanity and swearing. Posts with general profanity are not targeted, but they contain non-acceptable language.

Dataset :

The dataset is available at
https://hasocfire.github.io/hasoc/2020/dataset.html

Here the datasets for all three languages English , Hindi and German is available in a .tsv format. The table contains the unique_id , text, Task A result, Task B result, Task C result.

Approaches Taken :

We tend to implement three different approaches for the given task:

– Support Vector Machine without feature engineering

– Support Vector Machine with feature engineering

– Convolutional Neural Network (Optional)

SVM without feature engineering

In this approach, we first clean the given tweets by the following steps:

1. removing blank rows if any

2. replacing any digit with 0

3. modifying URLs to <urls>

4. changing text to lowercase

5. tokenizing each tweet into words

6. removing stop words and performing stemming

After the data is cleaned, the tweets are encoded as feature vectors and are directly used as input to our SVM.

SVM with feature engineering

In this approach, we will perform feature engineering before creating feature vectors for SVM. We select a handful of features which carry some relevant information about a tweet. For example, if the user uses one or more angry emoticons in a tweet, the tweet is more likely to be carrying the hatred emotion towards something.

We select the following features and include them in our feature vector:

– **emoticons:** we create a dictionary of happy, sad, anger, fear, surprise, disgust, others emoticons and count the number of each of them used in a tweet

– **hashtags**: we extract out the words used in hashtags in a tweet. Users often summarize their opinion through a hashtag. This can help find out the emotion expressed in a tweet.

– **intensifiers:** words like exceptionally, incredibly, awful, insanely, etc are often used to emphasize on some descriptive word. We call them intensifiers.

– **negations:** words like never, no. nothing, nowhere, etc are used to thwart the meaning of a piece of text.

– **hate words**: we use a list of hate words and find their occurrences in a tweet

We also hope to include other features like character n-grams, word n-grams, and few more..

Convolutional Neural Network (optional)

In the deep neural network approach, we will tokenize each input sentence and find the embedding of each word. Our hate speech dataset consists of tweets in three different languages. We will use domain-specific word embeddings which are trained on Twitter data.

We will pass the embeddings through our convolutional layers with multiple filter widths and feature maps. After each convolutional layer, we hope to perform max-overtime pooling before passing them through a final fully connected layer with output. We will train this model by minimizing the categorical cross-entropy loss.

Set up for the SVM and Neural Network:

For statistical machine learning-based approaches, we will use SVM with RBF kernel and c = 1.0 using grid-search and Random-forest with number of estimators = 50 (subject to change). We will use nltk3 libraries for all data processing steps in our SVM model.

For our deep learning-based approaches, we hope to use CNN. We will make our CNN deeper by using multiple filters - 3, 4, and 5. We will use different word embeddings for each language which we feed into our model as input.

– English: Used domain specific word embeddings trained on Twitter domain

– Hindi: Embeddings trained on Hindi corpus available in CFILT

– German: Embeddings from Europarl trained with FastText.

We will experiment with word embeddings of different dimensions to find the perfect fit for the problem.

Metrics for Evaluation:

Both recall and precision should be combined in the metrics for classification. There are many variants of the F1- score, such as weighted F1, Macro-F1 or micro-F1. The distribution of class labels for multi-class classifications is often unbalanced. The weighted F1-score independently calculates the F1 score for each class. It utilizes a weight dependent on the number of true labels of each class when it adds them. For the majority class, thus it provides a prejudice. For each class, the 'macro' separately calculates the F1 but does not use weights for the aggregation. If a system does not perform well for the minority classes, this results in a stronger penalization. The choice of the F1-measure variant depends on the task objective and the distribution of the label in the dataset. The classification issues linked to Hate Speech suffer from class imbalance. The F1 macro is therefore the natural choice for the assessment.