IRE Major Project Scope Document

SemEval19: OffensEval

Team 6

Sayan Ghosh, Aman Raj, Sahil Bakshi

1 Abstract

This paper describes our Task : OffensEval: Identifying and Categorizing Offensive Language in Social Media of SemEval-19. The project will consist of three tasks. First, identify whether the given post contains offensive language (Task1), followed by identifying whether the offensive language is targeted or non-targeted (Task 2) and finally identify the target of the offence- Individual, Group or Other. We show a Logistic Regression with Tf-Idf Model using which we achieved an F1-score of 76.5% and 51.2% 49.9% for sub-task A, B and C respectively and a SVM Model with 100 dimensional GloVe vectors using RBF kernel which we got 72%, 88% and 66% accuracy on Subtask A,B and C respectively. Through this document we provide a detailed description of the approach, as well as the results obtained for the task.

2 Methodologies Tried and Dataset Used

Dataset used: SemEval 2019 - OffenseEval Twitter dataset containing 13240 tweets.
Methodologies Implemented are as follows:

1. Logistic Regression Using Tf-Idf Vectorization And SVM with 100 dimensional GloVe vectors using RBF kernel

    - We implemented with Hierarchical(By taking successive layer as filter of input the 2nd level) and without Hierarchical Approach(By Inputting whole dataset for both 2nd and 3rd Levels)
    - We calculated F1 scores with maximum twitter length and also with average twitter length calculated to be 22 by data analysis
    - We also applied class balancing techniques such as upsampling/oversampling
    - We tried with and without @User Tweets Mentions as a feature for the vector space and got different accuracy/F1 results
    - We thus performed feature vector dimension analysis by reducing/increasing feature space
    - We finally performed Error Analysis and presented their reports as to how/why they occurred along with scores

2. Intro to Vanilla RNN

    - Just a basic model which we will improve upon in next phase of project
    - It has a embedding layer with 75 embedding space which learns along with the model

3 Findings from the current implementation

We have gotten the following results with the SVM and Logistic Regression models:

1. Logistic Regression using tf-idf vectorization

    - Subtask A
        - We need to see whether the tweet is offensive or not,It can be clearly seen that there in imbalance class that is number offensive tweets is 4400 and number of tweets not offensive is 8840,clearly there is class imbalance in the ratio 2:1.So on applying tf-idf along with normal data with logistic regression,we were getting 0.6763 f1score.
        - Average length-Earlier we have use logistic regression directly without seeing which tweet is longer or which tweet is shorter,to cope up with it we have now we have taken the tweets upto average length only which is 22 in given data sample set,although the feature space will be less ,but no major change was seen in f1 score which gives 0.66 ,same as earlier.

  - No. of mentions-@ is used to mention a user,but in earlier run with which we have 0.67 f1 score,we have not taken account of @,but know we have taken @user into consideration and we find our f1 score significantly dropped to
  - Class imbalance: Since the class ratio is imbalance ,we will use sampling to bring up the minority class to 1:1,on applying logistic regression along with tf-idf,we were getting f1score of .76,but although the f1 score has increased but it may results in over-fitting which is certainly not best for us.
  - Final Confusion Matrix Gives Result of [[2106 121],[670 413]]
- Subtask B
  - For hierarchical nature,we have first taken the tweets which are offensive and trained our model on it,and in non hierarchical we have not taken account of offensive tweets and we have trained model on full dataset.On testing we see hierarchical gives us f1score of 0.5 while non hierarchichal gives F1 score of 0.45, well even with good F1 score,we may find that hierarchical classification suffers with overfitting which is not good for us and also error propagation,we are propagating errors to last layer.
  - Using sampling to handle class imbalance we are bringing minority class in the ratio of 1:1 to major class,and upon testing the model,we get f1 score of 0.51 which is slightly improved but on larger datasets it may lead to overfitting.
  - Taking into consideration @user into account,we see f1 score around 0.46 but it is decreased just because giving

weight to @user in tf-idf is not the right way,there will be tweets with many @user that may not be offensive.

- Subtask C
  - On applying both the approaches we see the f1score comes around 0.48 ,but as we know hierarchichal has advantage that it will more precisely trained the day but also has some disadvantage that includes error propagation and overfitting is possible in hierarchichal.
  - We see if we take @user into consideration,f1 score comes around 0.47 although same but we have seen i earlier case that tweets having @user will not always be offensive.
  - We have again brought majority class in ratio with minority class,and f1 score comes around 0.48,but it will make model overfitted.

2. SVM with 100D pre-trained GloVe vectors with RBF kernel

   - Baseline accuracies with hierarchical approach and max length vectors -

Subtask	Test Accuracy	Precision	Recall	F1
A	0.72	0.64	0.43	0.51
B	0.88	0.88	0.99	0.93
C	0.65	0.70	0.84	0.79

- When hierarchical approach wasn't used, the accuracies for the subtasks B and C went down by around 10 percent each. Subtask A wasn't affected as it uses the entire dataset anyway.
- Hierarchical approach is better suited because we get rid of the extra tweets that won't be classified further and thus the training happens more efficiently with less noise.
- Using the max length of the tweet in the case of non deep learning clasifiers is better because the tweets don't need to be cropped to suit the average length of the tweets. Cropping may result in loss of important information. The extra 0s formed by taking max length for embeddings won't be that harmful to the classification.
- Hierarchical vs non hierarchical approach
  Subtask A - same
  Subtask B - hierarchical=0.89(training data score), 0.88(testing data score)
  non-hierarchical=0.78(training data score),0.68(testing data score)

Subtask C - hierarchical=0.83(training data score), 0.63(testing data score)
non-hierarchical=0.76(training data score), 0.61(testing data score)

- Data analysis with TP/FP values

Subtask	FN	TP	TN	FP
A	731	582	2361	298
B	0	1162	0	158
C	112	734	246	71

4 Code link to the baseline implementations

Github Link - OffenseEval: SemEval19

5 Error Analysis

My Model says it is offensive but the labels say its not: "thank u sm, this was worth pushing through the pain to be able to do things like that!!" -clearly mislabelled as offensive "he probably gets paid to say that..with $$ and assurances that he wont be called an islamophobe

–pffft its a fake word-jusin is an IDIOT and he is destroying canada one refugee at a time" -does not seem offensive

Strategies: 1.We use GLOVE,Word2Vec or fine-tuning on BERT. 2.Need of political awareness and quality control in creation of gold standard annotated by hand datasets

6 Methodology Difference and New Scope

We have implemented all the required methodology as mentioned in the project scope and deliverables. We have also performed the required Data/Error Analysis and the corresponding Scores/Accuracy. Also after looking into certain models have learnt that CNN might give better results than BERT and also NB will not be used anymore and will thus, continue with CNN and RNN/LSTM in future along with current Models.

7 New Timeline for final deliverable goals

The New Timeline remains the same as the Project Scope Document. 2nd November - Mid-Evaluation for Project and Discussion with Mentor on the RNN/LSTM and CNN Models. 10th November - Project Completion and Detailed Presentation.