IRE

2nd Deliverable - IRE Major Project

SemEval19 - haTEval

Tanuj Garg(20171106), Mashrukh Islam (20161137)
Saurabh Chand (20161106), Alok Kumar Kar (20171103)

Abstract

Hate Speech is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics. With the rise of social media and user generated content, detecting and classifying hate speech is becoming quite important. All platforms do have some form of manual checking for this but it can't be as fast as automated classification.

We will be identifying if the target of hate is a single human or a group of persons and if the message author intends to be aggressive, harmful, or even to incite, in various forms, to violent acts against the target, specifically on social media platforms.

1. Task 1: Classify hateful tweets (where Hate Speech against women or immigrants has been identified) as aggressive or not aggressive. Note that aggressiveness is not a mandatory characteristic for all hateful texts and some text can express hate against a target in terms of disrespect without using an aggressive language.
2. Task 2: Classify hateful tweets to identify if the target harrased is a generic group of people or a specific individual.

The dataset is a twitter dataset which was provided to us. It contains 9000 training set tweets and 1000 validation set tweets.

Methodologies

Preprocessing: The model takes the given string data and pre processes it, adding 'USER' for '@' and replacing URLs with 'URL'. Removing emojis. And finally stemming the tokens. The last part is necessary because the dimensionality of the feature set would be much higher because the different types of the same words would represent different features.

Vectorization: We're using scikit-learn library Count-Vectorizer to create feature-vectors from our data and we're only using word-level n-grams for this. We pass this to tf-idf transform to get tf-idf vectors.

Final Predictor: In the end we use either a Multinomial Naive Bayes model, Logistic regression model or SVM model.

Naive Bayes Model

Running this model on the test set given we get the following result.

	Accuracy	Precision	Recall	F1-score
Task 1	0.6288	0.6381	0.6212	0.6136
Task 2	0.8605	0.8709	0.8623	0.8599

Logistic Regression Model

Running this model on the test set given we get the following result.

	Accuracy	Precision	Recall	F1-score
Task 1	0.6323	0.6320	0.6292	0.6287
Task 2	0.8829	0.8862	0.8842	0.8828

SVM Model

Running this model on the test set given we get the following result.

	Accuracy	Precision	Recall	F1-score
Task 1	0.697	0.798	0.5737	0.5375
Task 2	0.71	0.70	0.739	0.66

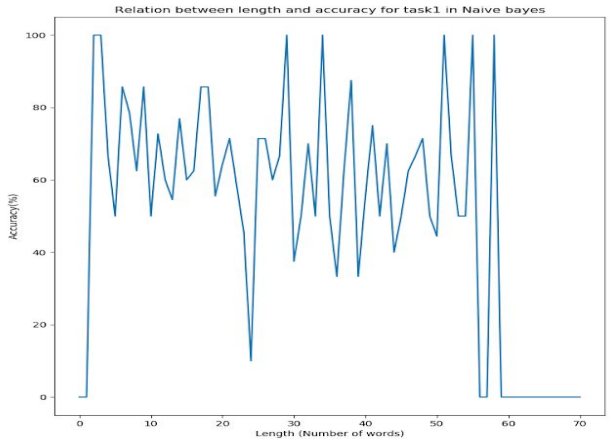
Findings:

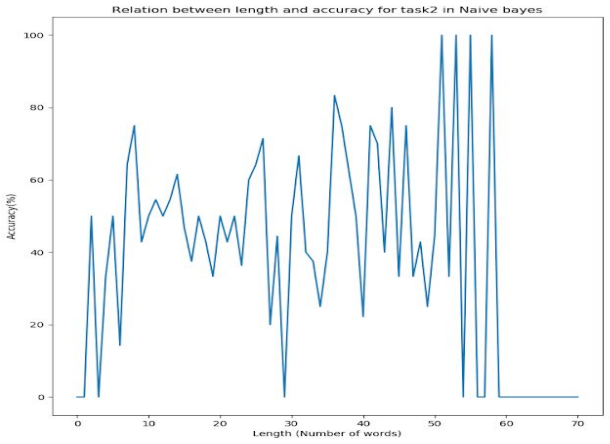
We have twitter data, on which we perform some analysis.
Maximum tweet length: 63
Average tweet length: 22.12
Minimum tweet length: 1
Percent tweets with lengths less than avg: 58.3
Number of hateful tweets: 3783
Number of non-hateful tweets: 5217
Number of individual targeting tweets: 1341
Number of group targeting tweets: 7659
Number of Aggressive tweets: 1559

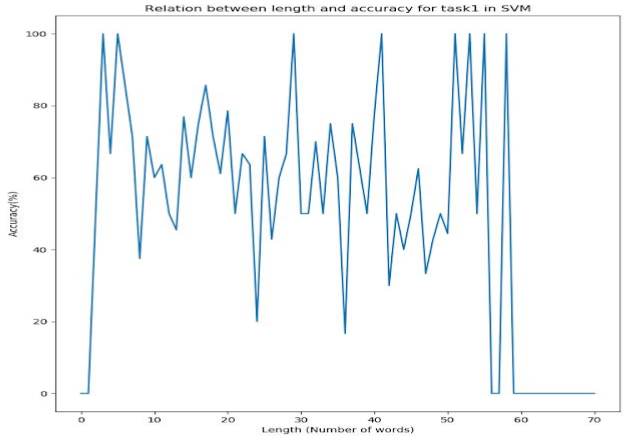
Number of Non-Aggressive tweets: 7441
Checking to see if all the aggressive tweets are hateful:

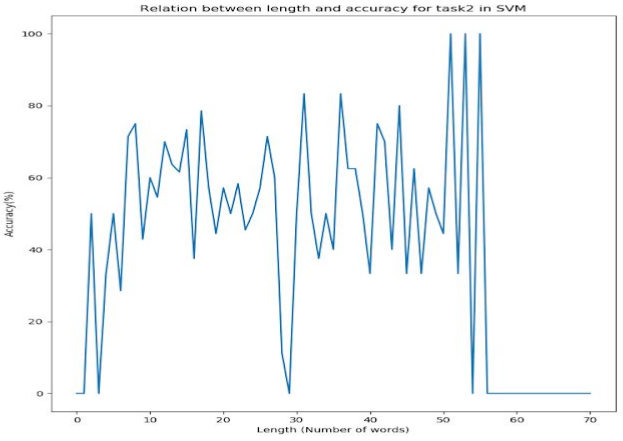
	Hateful	Non-Hateful
Aggressive	1559	0
Non-Aggressive	2224	5217

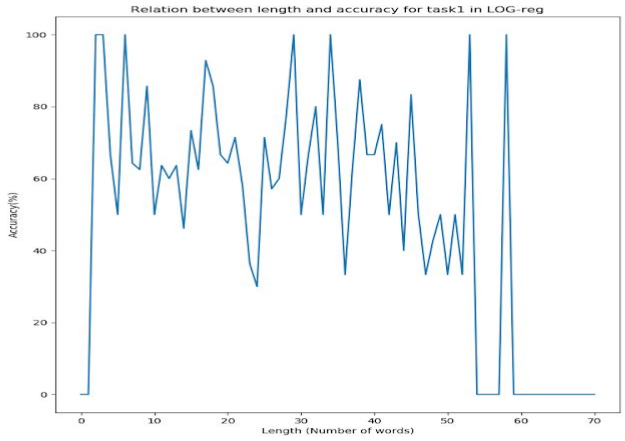
	Hateful	Non-Hateful
Individual	1341	0
Group	2442	5217

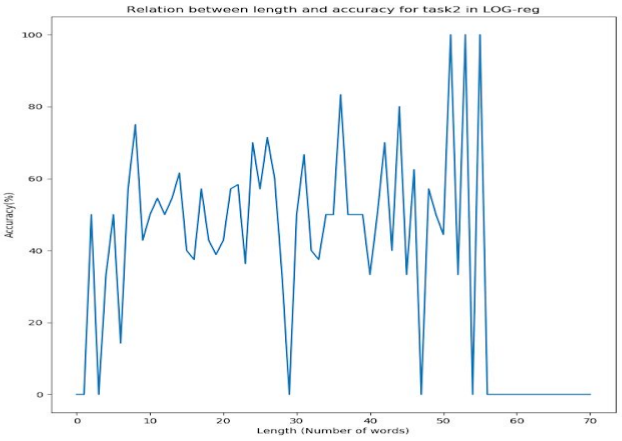












These are the correlation graphs between the lengths of tweets and accuracy of our models. We can see that for task 1 SVM performs better than the other models, however underperforms in task 2.

Comparison to Original Timeline:

We have achieved aggression and target identification using models to identify target and aggressive tweets as per the original scope document. The methodology used and the implementation of baseline was as specified in the scope document. We have implemented the baseline code as per the original timeline provided.

CODE LINK : https://github.com/alokkar/ire-major-project-group-8

Final Deliverable Timeline:

One of the main problems is having a small dataset. We plan to overcome that by using data augmentation, where we translate the tweet to another language and then back to english where the sentiment would remain the same for most cases but the wordings would change adding more information to the data present.

We also plan to see if sequence models (like CNN or LSTMs) perform better on this or not. We also plan to do a bit more of data analysis like effect of emojis and user mentions on our models.

Another note is the removal of stop words. It might be beneficial to keep the stopwords or remove them. Their effect in the models would be checked later and the better alternative would be chosen.