CS 6350: Project Milestone 2

Name: Tejas Ramesh Pawar — u1452462

April 4, 2024

- **Updates after first milestone**

  In the previous milestone, I implemented a simple perceptron on all the three datasets and used the *tfidf* dataset for my kaggle submission, since it gave me the highest accuracy.

  This time I also used the *miscelleneous* dataset along with these datasets. What I did is I appended the *misc* dataset with these three datasets. In order to append the dataset, I had to do some cleaning on it.

  The cleaning procedure: I replaced the *defendant_gender* column with *male*, *female* and *indeterminate* columns, and gave the values as 0, 1 accordingly. Similarly, I did the same with *victim_genders*.

  I did the same with *offence_category* feature, where I numbered each of the offenses from 1 to n(number of unique offenses).

  This enriched the dataset, and when I ran this on the dataset, it increased the accuracy. The last accuracy I got on the test dataset was 69%, but using this I got an accuracy of 83.15%. The kaggle submission with a score of **0.81** is the same. I tried the decay perceptron and the averaged perceptron on these datasets as well, but did not see any improvements.

  The next step I did was to use a new algorithm, i.e. Support Vector Machine(SVM)(without using the libraries). I implmented this algorithm and used a 5-fold cross validation on it and found the best *learning rate* and *lambda parameter*. I ran the same with 100 epochs and got a test accuracy of 80.75%.

  I used this algorithm on all the three datasets with the *misc* dataset appended to each one. Here too, the *tfidf* performed the best. The kaggle subimission for this has a score of **0.80**

  Another thing that I did which does not count to my final kaggle submissions or won't count towards my final code submission(probably) - I appended the *tfidf*, *bow* and *glove* dataset one by one in combinations to check which performs the best. But none of the combinations excelled.

- **Plan for the rest of the semester**

  I will be using other algorithms - Logistic Regression and Ensembles using mixture of algorithms above algorithms I implemented.

- **Challenges faced**

  I faced some challenges while implementing the SVM alfgorithm, which I resolved fairly quickly. The only major challenge was the trial and error I did combinining different datasets in different combinations, which was time consuming and tedious.