**Sentiment Analysis Report**

Based on the problem statement the goal is to predict the sentiment towards the named entity. I have followed the below approach and implemented technicalities which will suffice the problem statement.

As a preliminary requirement, I have imported some libraries and used some validation techniques which will help in loading, cleaning, analysing, partitioning, building, and validating the mode like pandas, NumPy, sklearn, TensorFlow, Nltk.

As a first step of my exploratory data analysis, I have loaded the dataset and verified the size of the dataset, basically the number of rows and columns. Post that i have verified the datatypes of each attribute and then i have checked for discrepancies like any null values in dataset and checked for duplicates as well, post identification of duplicates i have removed the duplicates as having duplicates will just increase the noise in the data.

To analyse each attribute, I firstly performed univariate analysis on column ‘Entity’ and identified 748 types of categories in Column ‘Entity’. Out of 748 categories there are 333 categories having frequency as just 1. So, I have converted all these categories into Others.

As we cannot consider data in string format for predicting sentiment, with the help of one hot encoding I have converted all 416 categories of column ‘Entity’ to separate columns having binary values as 0 and 1 where 0 indicates the category presence as ‘No’ and 1 indicates ‘Yes’.

To handle unknown categories/vocabulary in entity columns where our model was not trained, I have included a function to ignore the unknown categories we may receive in future. The reason for doing this activity is just to avoid any errors which our model may encounter if it sees any unknown categories.

As part of cleaning data, I have identified and removed some stop words and removed some special characters in column ‘Sentence’ which do not contribute in any way to my prediction.

I have applied TFIDF vectorization to just store the frequency and repetition of each word in a dataset and post that I have transformed and saved in ‘sentence Col’ which I will use in model building phases.

For the Building model, I need to separate independent(X) and dependent variables(Y) and post that I have divided data into training split and testing split.

As the first part of my model building, I have created a function which will include algorithms like Logistic Regression, Support vector machines, Gaussian Naive Bayes and Stochastic Gradient Descent. I have included a cross validation technique as well just to ensure whether my model is performing consistently on any kind of data it receives and I have split the data into 5 folds as part of my cross-validation process. Then I have added validation/Evaluation methods like precision, recall, F1 Score and I have developed the function in such a way where it will provide model wise performance in a single data frame format which is easily readable/Interpretable.

Post building and evaluating all the models, I have tried with Artificial Neural Network as well and results are not encouraging as I can see it’s over fitted.

Post evaluating and comparing all model’s performance, my stochastic Gradient descent algorithm has given good results and has performed reasonably well compared to other algorithms. I have taken F1 score as my metric to evaluate the model performance.

I have pickled my model and post that I have repeated the entire pre-processing on test data as well and extracted predictions on test data as well.