Student name : Loïc XU

Title : Sentiment Analysis on Airline Tweets

Theme : In this project, I aim to perform sentiment analysis on tweets related to airline experiences. So, basically try to predict the sentiment felt by the author of the tweet about US Airlines, 3 possible cases : positive, neutral, negative (multiclass classification). This idea came from the fact that recently, a door of Alaska Boeing 737 plane was ripped-off in mid-air and passengers were tweeting on the event and posting videos.

Dataset : Twitter US Airline Sentiment in 2015 from https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment/data?select=Tweets.csv

I plan to use a Kaggle dataset containing tweets about various airlines, along with their associated sentiments (positive, negative, or neutral). This one stands out from the other on Kaggle as it is said to be human labelled which is important to avoid impurity in the dataset which will affect the performance of my models. The dataset contains 14640 entries and has 15 columns. The ones that we are interested in are "airline_sentiment" and "text". I do not think that the other columns are relevant in our context as we do not want the model to overfit on for instance Twitter's username or airline name.

Algorithms/Methods : First, I will preprocess the data as seen in the labs with tokenization, lemmatization/stemming, removing stopwords and removing the name of the airline in the tweet as each tweet starts with “@Name_airline”. Then, I will try to do feature engineering and use techniques like TF-IDF to find the words that are not too frequent but not too rare or also N-words.

For the algorithms, I plan to use :

- Naïve Bayes, it will be our baseline and the other models should try to beat it.
- Softmax Regression, same as logistic regression for more than 2 classes.
- Random Forests
- CNN / RNN, deep learning is the new buzzword, I should try to implement one too for this project.
- Fine-tuning pretrained models, utilize a pre-trained language model such as BERT to perform transfer learning on the sentiment analysis task. Fine-tune the model on the airline tweets dataset to adapt it to the specific context of airline sentiment analysis. This will serve as a benchmark for comparing the performance of my models against state-of-the-art (SOTA) models.

- … Maybe more ? If I find other promising models, I will include them as well.

I will most likely not implement from scratch these models and use libraries already made. Except maybe for some that are not too long to code.

Evaluation : For the Evaluation, I will use the classic metrics as accuracy, precision, recall, F1-score with the splitting method of train/validation dataset. I will also compare all the models between each other. And finally, I will try on some recent tweets about the Alaska Airlines and its Boeing 737 and other tweets that I will try to find related to Airline.