* **-VIVEK KUMAR**

**Problem Statement:**

Create an API endpoint that can accept a text and return associated sentiment with it.

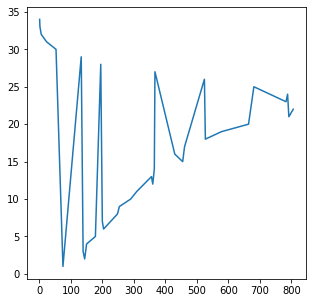
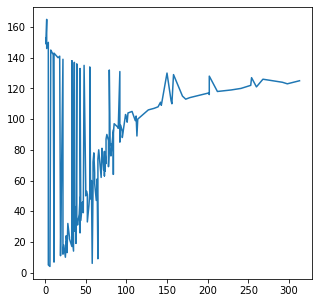
**Experiment setup:**

Platform: Colab

Device: Gpu

Database: Drive

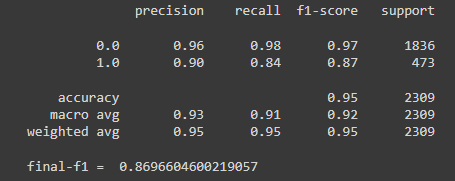
**Data Insight:**

1. The first take away from the dataset is that it includes the customer experience regarding the different airline (majority negative and mostly complaints).
2. Dataset comprises of mostly negative sentiments, Highly Skewed.
3. Next, I checked the length trend in the text, it appears that most of them are under the truncation range for most of the models ~150.
4. Max length of sentence in the text is of 34 words.
5. The dataset is comprised with the data of 6 different airlines namely virginAmerica, united, southwestAir, Jetblue, USAirways & AmericanAir.

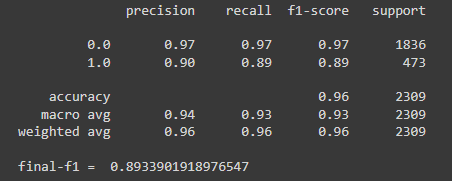
**Solution Approach:**

1. Since the dataset was highly biased, my naïve approch was that, to use **RandomOverSampler** to remove bias towards the majority class.
2. But for the mainstream model approach, I used **Matthews correlation coefficient (MCC)** for binary classification evaluation as it is an alternative measure unaffected by the unbalanced datasets issues.
3. After preprocessing the dataset, I’ve tested 2 model architectures BERT and RoBERTa by finetuning them on train dataset (mydf\_dataframe).

* BERT had a score



* RoBERTa had a score



1. I have used **distilRoBERTa** since distilled versions of the transformers models compromises with few accuracy points for the sake of quicker training and inference.
2. Reason for using the **XLNet** model, as it compromises with speed of training and inference in exchange for potentially better performance on complex tasks.

**Ablation study table:**

| **Concept** | **Fine**  **tunning** | **Model**  **name** | **MCC**  **Score** | **F1 Score** |
| --- | --- | --- | --- | --- |
|  | Yes | bert-base-  cased | 0.838 | 0.869 |
|  | **Yes** | **roberta-**  **base** | **0.866** | **0.893** |
| **Binary Classification**  **(Sentiment Analysis)** | Yes | xlnet-base-cased | 0.841 | 0.872 |
|  | Yes | electra-base-discriminator | 0.860 | 0.887 |
|  | Yes | distilroberta-  base | 0.845 | 0.875 |

**Conclusion:**

* In this experiment, **RoBERTa** seems to outperform the other models with overall **86.6** performance.
* I would have hyper tunned the **ELECTRA** model a bit more before trying to finalize any approach for production. Since ELECTRA can make up the difference, according to the current GLUE benchmark [leaderboard](https://gluebenchmark.com/leaderboard/), it is sitting above RoBERTa.
* I also came across a nice paper on **Aspect-based Sentiment Analysis** **(**[**ABSA**](https://www.cs.uic.edu/~lzhang3/paper/ZhangLiu-AEEE.pdf)**)** which would be very useful since it seeks to extract the most essential elements of an entity from text and predict their polarity. A review of ABSA’s most recent SOTA reveals a significant increase in discovering both **aspect and sentiment**.



* This assignment was enclosed with the various NLP concepts and addresses the numerous aspects and use case of them

**References:**

* <https://www.cs.uic.edu/~lzhang3/paper/ZhangLiu-AEEE.pdf>
* [https://bmcgenomics.biomedcentral.com/articles](https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6413-7)

**Assignment links:**

* **Google Collaboratory:-** [Colab folder and codes](https://drive.google.com/drive/folders/17I38S6HoFZch7D_TSQqHHBGNKonkWb2u?usp=share_link)
* **Model deployment on: -**

1. **Hugging face: -** [velvrix hugging model](https://huggingface.co/velvrix/truefoundary_sentimental_RoBERTa)
2. **Streamlit: -**  [velvrix streamlit](https://velvrix-streamlit-cloud-app-f0lmlu.streamlit.app/)

* **Github links: -**

1. [Truefoundary\_fastAPI](https://github.com/VELVRIX/truefoundary_fastAPI)
2. [Truefoundary\_streamlit\_cloud](https://github.com/VELVRIX/STREAMLIT_CLOUD)