Free Topic : Sentiment Analysis on Real-Time Streaming Tweets

Team Ninja
Team Members: **Captain-asankar2@**

1) Kunika Sood: kunikas2@
2) Arvind Sankar: asankar2@

What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?

Our Free topic is Sentiment Analysis. Sentiment Analysis is used to identify the sentiments of the text source. Very commonly, Tweets are used in collecting lots of data for performing sentiment analysis. They can help us understand the opinions and perspectives of various people about a particular topic. The task is to capture the sentiments of the data by developing a ML pipeline and then evaluate the results of the model.

We will use the Twitter Historical Tweet APIs [1] to provide a dataset of tweets on a show like “Squid Game” for a definitive timeline. After collecting a substantial dataset, we will hand label the tweet data and mark it as positive or negative sentiment. We can then do analysis and preprocess the tweets to prepare for a specific model. The model created can be validated and then evaluated with F1 Score (precision and recall).

We will evaluate the data on streaming data via the Twitter Streaming APIs [2]. As tweets for a topic appear in real time, the model will evaluate each tweet as positive or negative. After the tweets stream, we can hand label them and evaluate the accuracy of our model. If we choose to evaluate multiple models, we can measure evaluation speed as another metric.

This task is interesting since it could allow people to measure audience sentiment in real time. If a show is airing, producers of the show can use this tool to understand what the audience thinks of each scene. This would give them a level of insight that would make subsequent episodes more engaging for the audience.

Dataset: We will be using Twitter data for the project
Tools: For streaming, scrapping and stemming Python can be used for data in real time.
Evaluation: We will evaluate using Precision/Recall measures.

Which programming language do you plan to use?

We are planning to use Python.

Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

Justification of Workload: 20 X 2 = 40 hours

Phase	Task	Estimated Time
1. Collect historic tweet data	Create Twitter developer account, learn which APIs we need, etc.	~5 hours
2. Fetch tweets relevant to topic at a defined time interval	Write code that uses API to fetch tweets relevant to a pre defined topic at a predefined time interval	~2 hours
3. Data Preprocessing	Label Training Data	~3 hours
4. Data Analysis and Preprocessing to Prepare for Specific Model	Apply stemming, tokenization, stopword removal, etc.	~8 hours
5. Split data as Training and Test Subset	Split labeled data as Training and Test set in N samples (70/30 ratio)	~1 hour
6. Model Creation with Training	Train models	~ 5 hours
7. Model Validation with Test Data	Validate the models with test data	~ 6 hours
8. Run Experiments on Streaming data	Write code that fetches data / applies preprocessing / sends to the model / records result	~5 hours
9. Evaluate the results of the model on Streaming data.	Label Testing Data	~1 hour
10. Presentation and demo video	Record a demo of model running evaluations on a live twitter topic.	~4 hours
Total		~40 hours