Progress Report: Sentiment Analysis on Real-Time Streaming Tweets

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

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

Tasks that have been completed:

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

We have written code to download tweets and metadata of tweets and have built a small dataset for our models to train on.

Which tasks are pending?

Started performing various pre-processing steps on the dataset to remove urls, punctuations and converting data to lowercase for better generalization.

Task	Details	Total Hours	Hours used
1. Data Preprocessing	Label Training Data	~3 hours	
2. Data Analysis and Preprocessing to Prepare for Specific Model	Apply stemming, tokenization, stopword removal, etc.	~8 hours	~2 hours
3. Split data as Training and Test Subset	Split labeled data as Training and Test set in N samples (70/30 ratio)	~1 hour	
4. Model Creation with Training	Train models	~ 5 hours	
5. Model Validation with	Validate the models with test data	~ 6 hours	

Test Data			
6. **Run Experiments on Streaming data**	Write code that fetches data / applies preprocessing / sends to the model / records result	~5 hours	
7. **Evaluate the results of the model on Streaming data.**	Label Testing Data	~1 hour	
8. **Presentation and demo video**	Record a demo of model running evaluations on a live twitter topic.	~4 hours	

Are you facing any challenges?

One challenge we have faced is the limitation of the Twitter API. Since we are using the developer account to retrieve tweets, we are only able to retrieve tweets from the last 7 days.

This presents an issue since we were planning on training models on TV shows that air each week, but with this approach we would only get one week's worth of data. Instead, we pivoted to training on a TV show that airs daily (Jeopardy) to collect sufficient data.

We have already collected one weeks worth of data with our current tooling. We can collect another week's worth of data starting next Saturday.