Assignment 3

Submitted by:

Abirami Dhayalan - axd210002
Harish Srinivasan - hxs200044

Question 1:

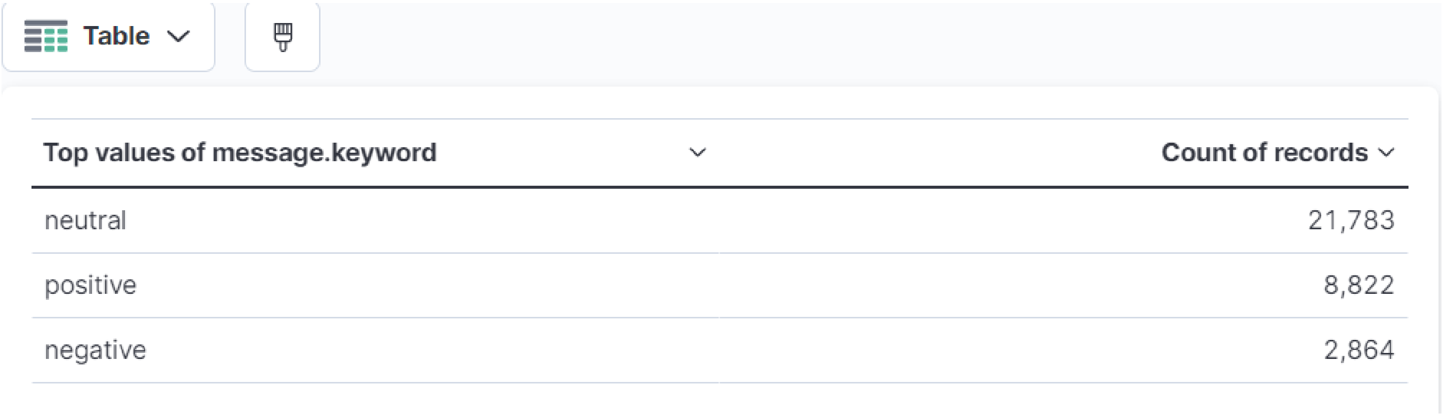
1. **Overall workflow:**

Twitter API -> Tweepy/TCP Socket -> Spark streaming (Sentiment Analysis) -> Kafka topic -> ELK

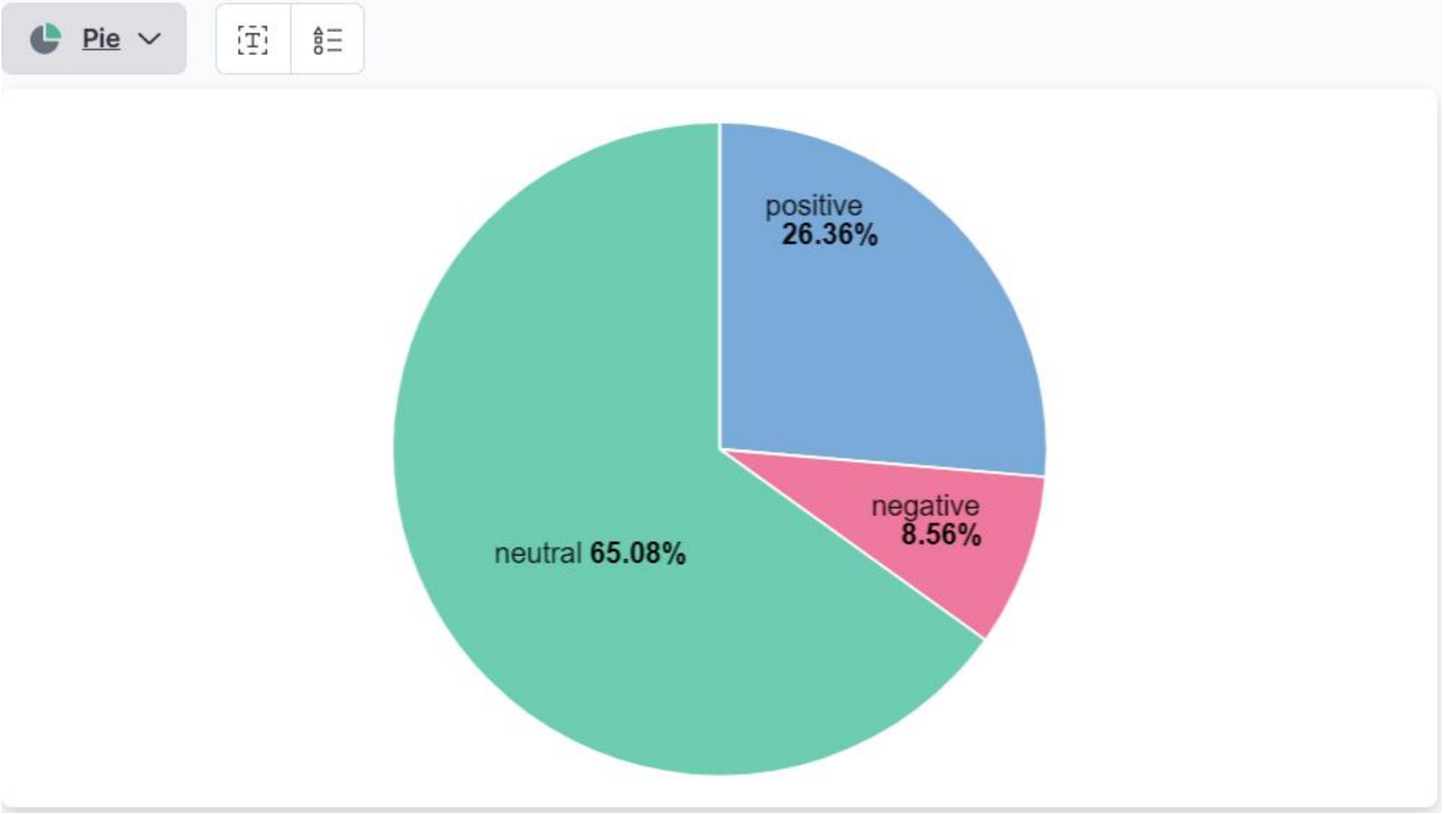
2. **Execution Steps:**

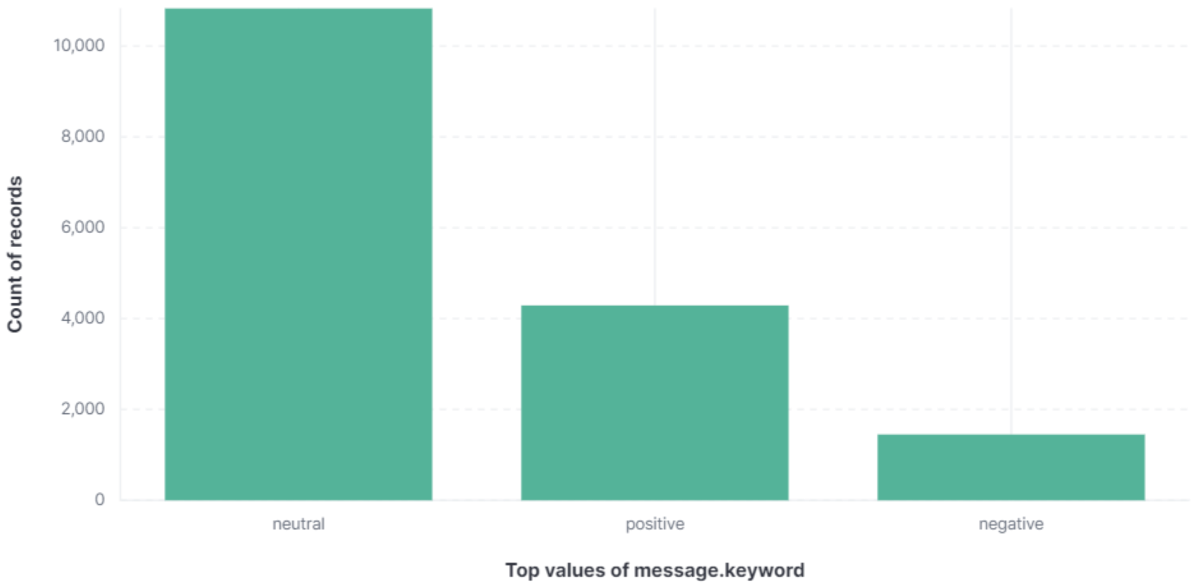
- Step 1: Start Zookeeper and Kafka services.
- Step 2: Start elasticsearch, kibana, kafka server, logstach services
- Script1 - Using tweepy library to collect tweets through tcp socket.
- Step 3: Then create a kafka topic.
  .\kafka-topics.bat --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic sentiments
- Script2 - Listens to the socket and preprocesses the tweets. The sentiment classifier model classifies the sentiments of the tweets received as positive, neutral and negative. The sentiment outcome for every batch of tweets is pushed to kafka topics 'sentiments'.
- Step 4: Integrate kafka and elk, Configure logstack - input: kafka topic containing sentiments and output is elastic search index 'sentiments'. Then the constant stream of data will be available for visualiztion in kibana
- Script3 - configuring logstach, we use the tweets from kafka topic and send them to elastic search

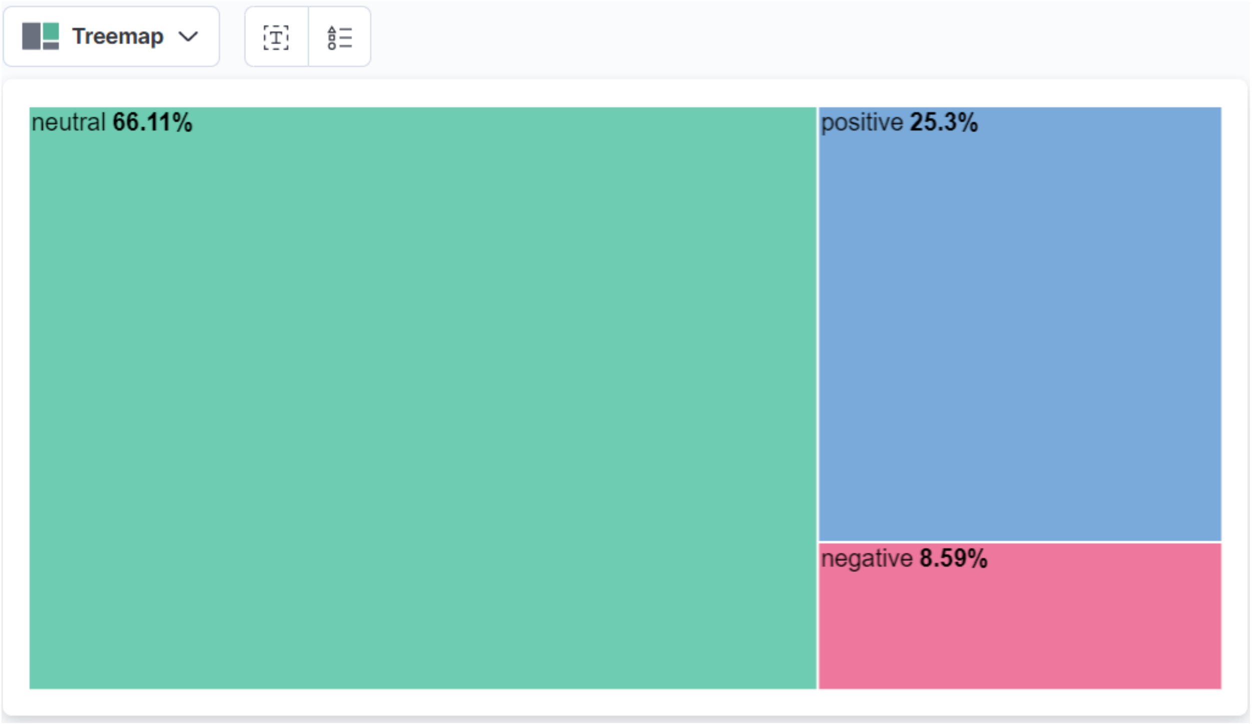
3. Visualizations:



Top values of message.keyword	Count of records
neutral	21,783
positive	8,822
negative	2,864







4. Insights:

- In order to receive more tweets, we decided to pick a popular trending topic so the tweet hashtag – “Russia” was chosen.
- Overwhelming majority of the tweets are neutral. This may be due to the fact that most tweets are news stories reporting about Russia.
- Minor fluctuations in plots were seen since there was an incoming stream of tweet data which was dynamically varying.