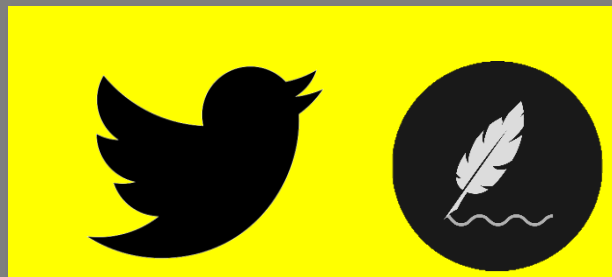
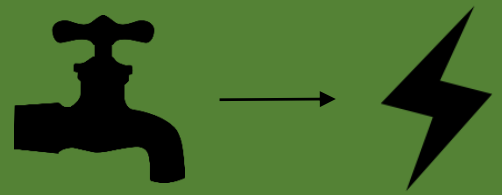


1 The Twitter Story

2 Why Sentiment Analysis?

3 Approach



4 Topology - Storm

5 Visualization

6 Scalability

Robin Singh Nanda
S/R - 11022

Nithin Raju D
S/R - 12393

History

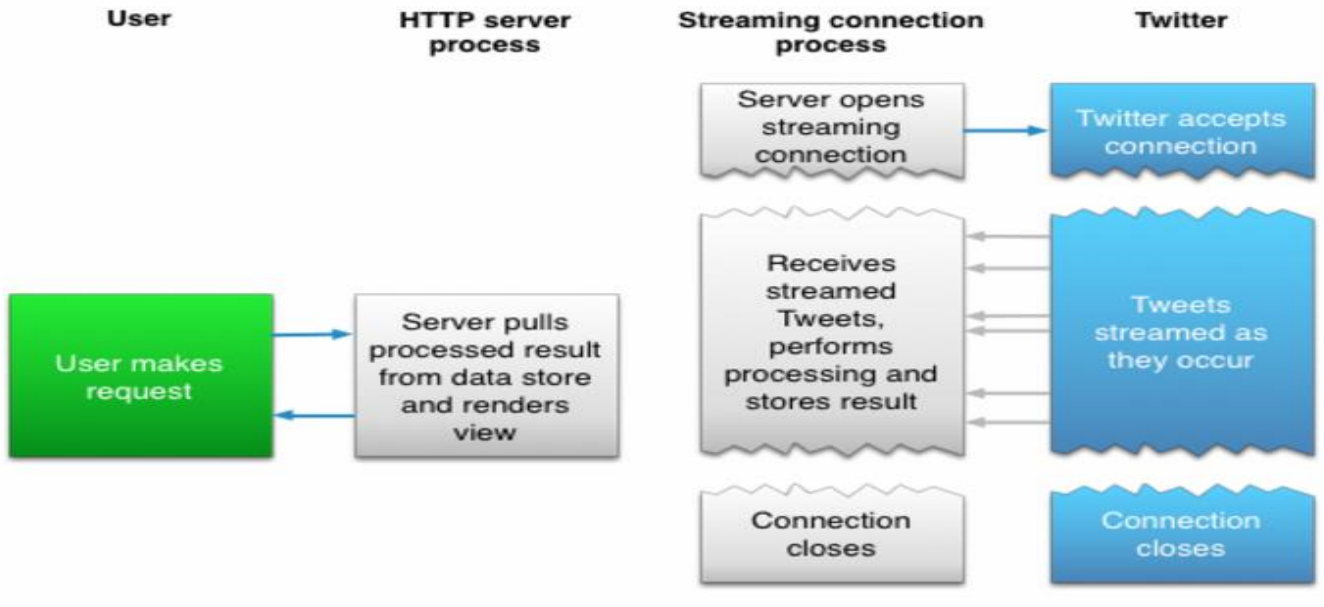
EST. 2006

- Jack Dorsey

- Online social networking service

- The company started with the idea of an individual using an SMS service to communicate with a small group.
- Short messages (140 characters) called ‘tweets’. Experienced rapid growth from 400,000 tweets per quarter in 2007 to 100 million per day in 2010.
- Concept of followers where other users can follow a person to understand latest updates about their actions.

Twitter Streaming API



Credit: Twitter Streaming APIs overview

- Low Latency access.
- Several streaming endpoints.

- Parsing, filtering and aggregation of data possible.
- Unlike Firehose, publicly available data - limited access (~1% of random sample).

Sentiment Analysis

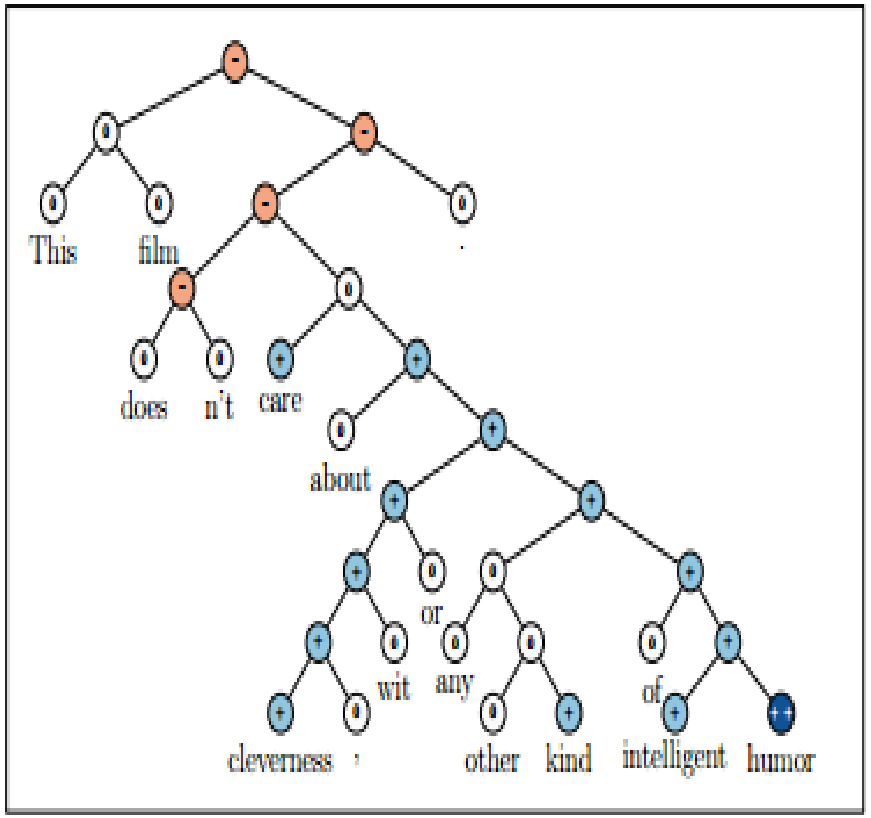
- Sentiment analysis – otherwise known as opinion mining is the process of determining the emotional tone expressed within an online mention.
- Helps to gain an overview of wider public opinion.
  Ex: Obama administration used it during 2012 presidential election to gauge public opinion
- Few use cases - Brand image, why consumers didn't buy the product? Understand gaps in existing offerings?

Approach

❶ Emoticon ☺ ☹ 😐

- Polarity of word based on the emoticon in tweet
- Naïve Bayes Classifier for sentiment analysis
- Constraints
  - Data limited due to filtering
  - Meaning of sentence not considered

Approach

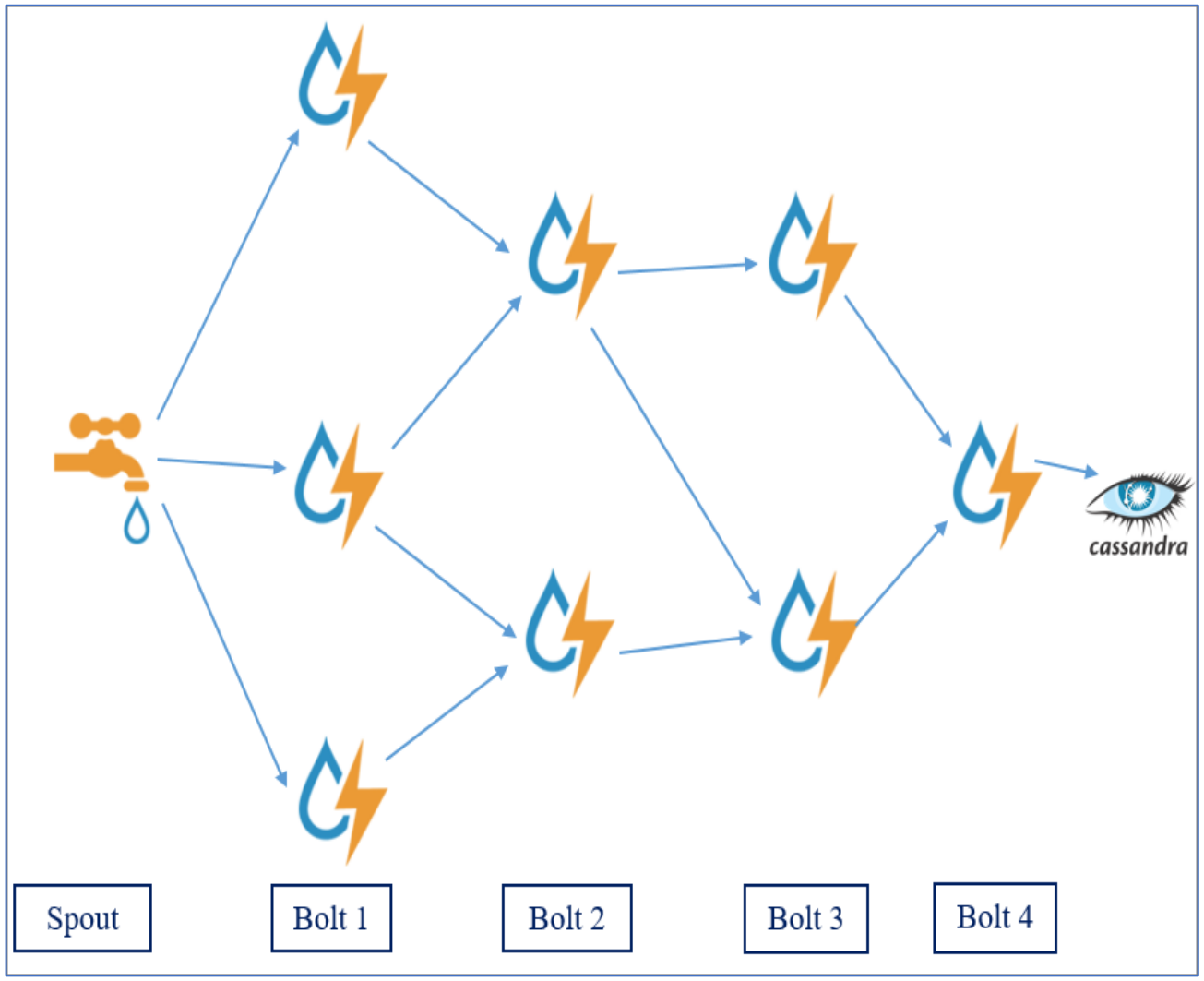


Recursive Neural Tensor Network example predicting the sentiment at each node of the parse tree

2 Deep Learning NLP Model

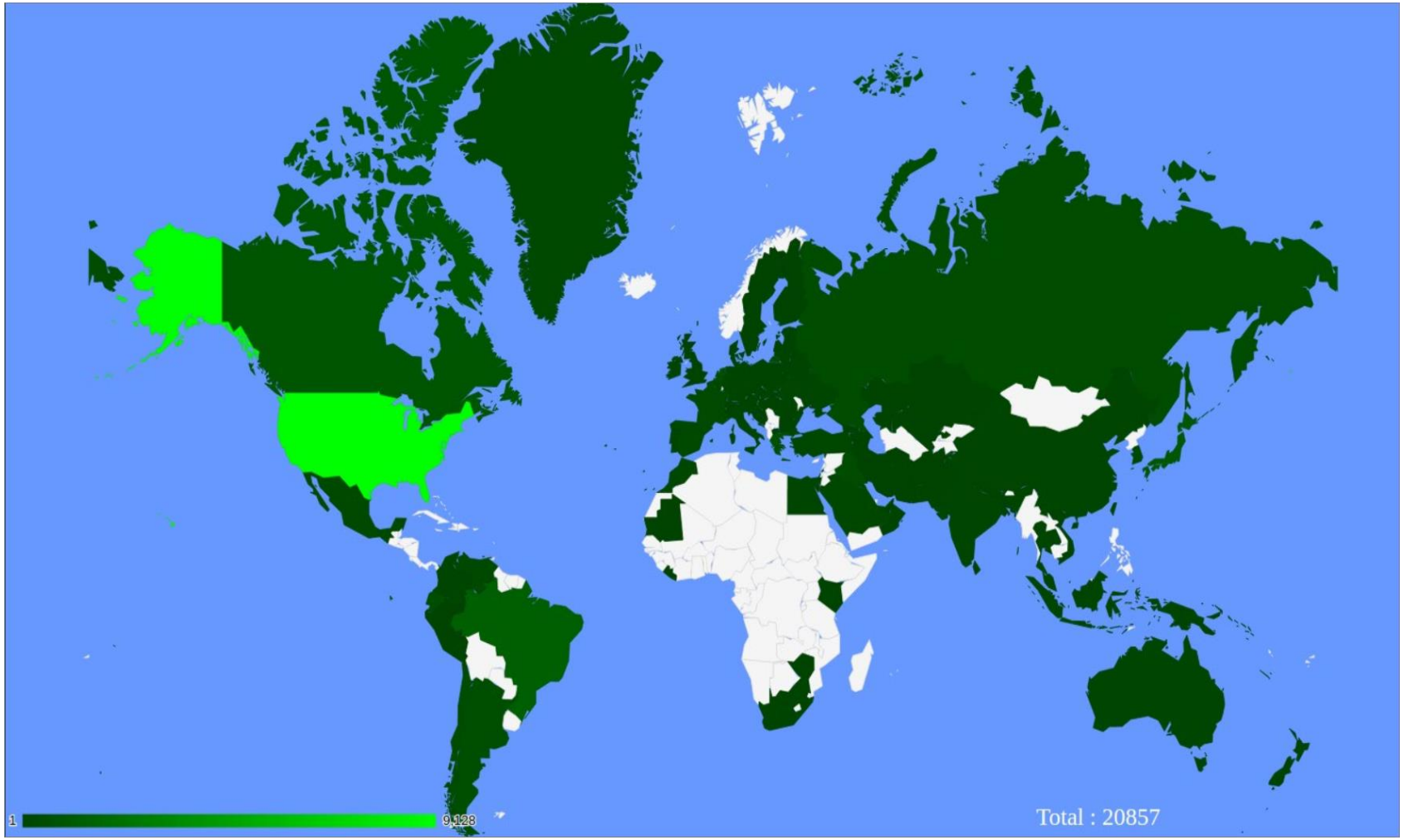
- Sentiment based on the sentence structure
- Stanford Sentiment Treebank & Recursive Neural Tensor Network
- Better sentiment classification
- Twitter4j and NLP java libraries used

Storm Topology



	INPUT	OUTPUT
SPOUT	HDFS File/Twitter Streaming API	Tweets + other related data such as time zone, user details, media, geolocation etc.
BOLT 1	Tweets + other related data such as time zone, user details, media, geolocation etc.	Sentiment (Positive, Negative, Neutral) & Time zone of tweet
BOLT 2	Sentiment (Positive, Negative, Neutral) & Time zone of tweet	Sentiment & Country mapped to time zone
BOLT 3	Sentiment & Country mapped to time zone	Count of positive, negative, neutral sentiment (By country) & country
BOLT 4	Count of positive, negative, neutral sentiment (By country) & country	Store to Cassandra DB

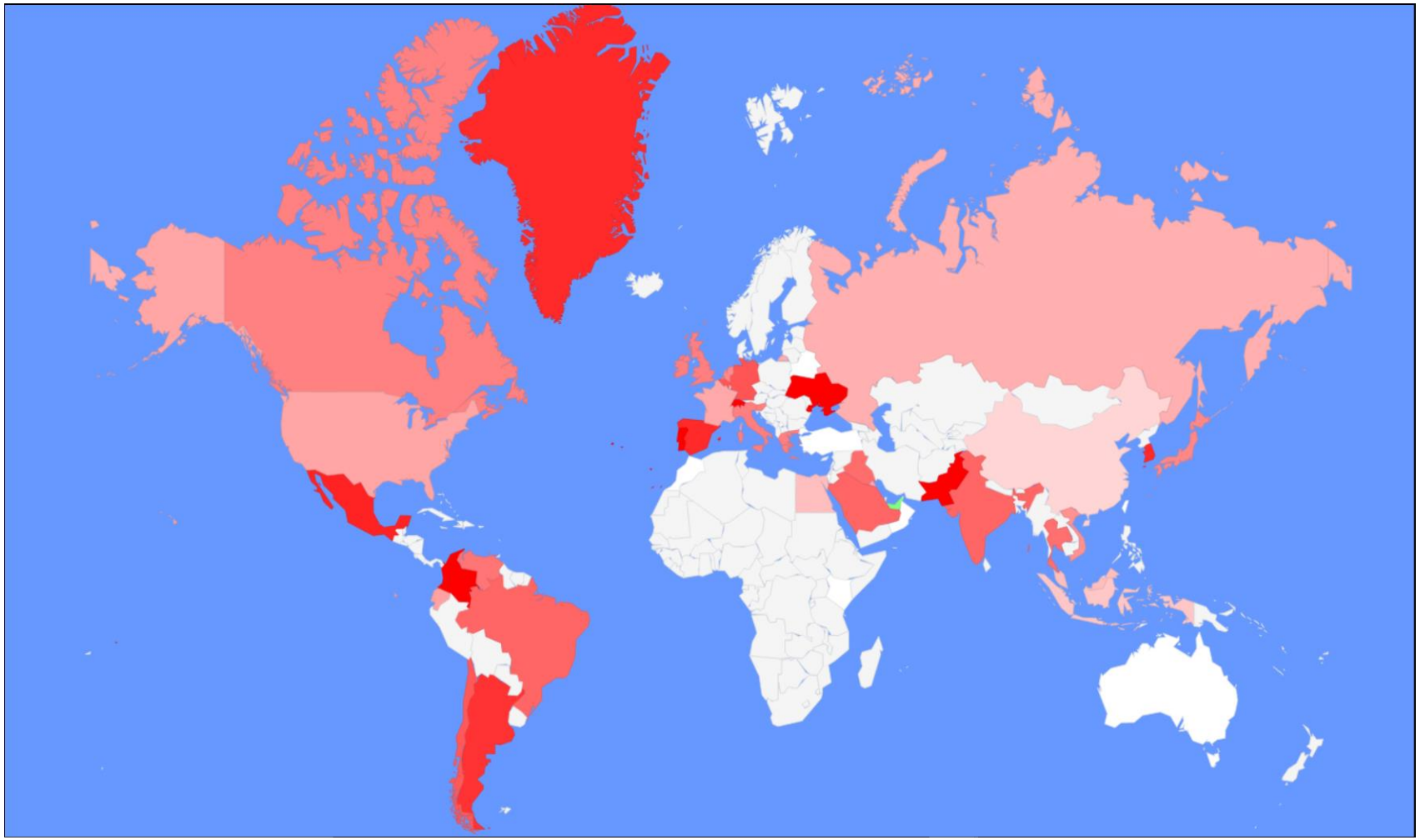
Visualization



Count of tweets in the world (Twitter Streaming API)

- Code was run for 10 minutes.
- Tweets streamed: 20,857
- More number of tweets in US and other western countries
- In Africa, a negligible twitter presence observed.
  - Power Shortage
  - Internet reach
- Information asymmetry potentially more in developing/poor countries.

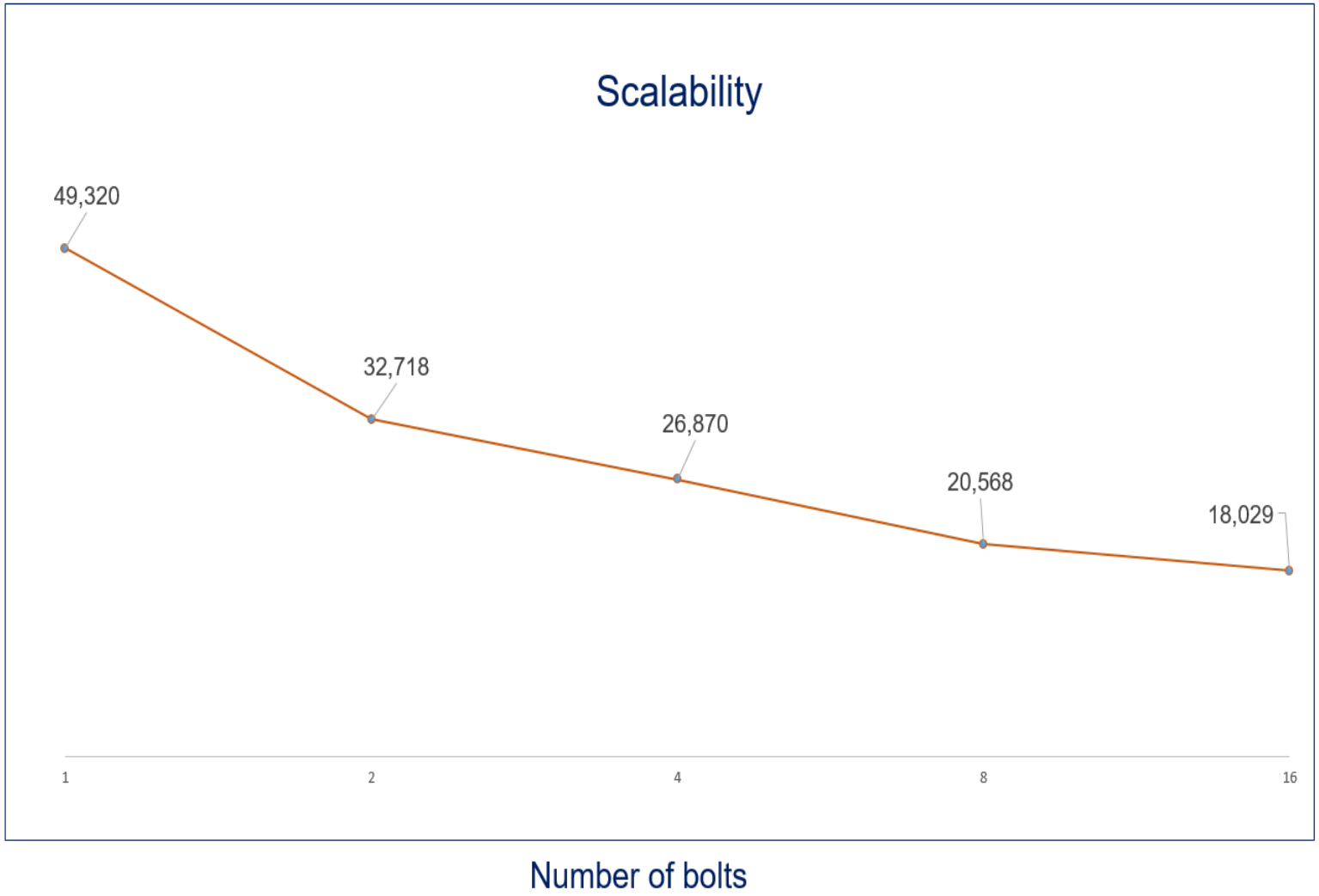
Visualization



Sentiment of tweets in the world

- Code was run for 20 minutes.
- Tweets streamed: 45,675
- Few parts of US and South America and India show higher negative sentiment relatively
- Visualization in sync with previous. Sentiment is neutral in Africa potentially due to very less Twitter presence

Scalability



- Rate at which tweets were streamed were increased
- Number of bolts increased to handle the increasing stream rate
- Decreasing trend
- Execution time almost stabilizes around 18 to 17 s

LIVE DEMO

References

1. Internet live stats | Twitter Usage Statistics
2. Twitter Usage | Company Facts
3. Twitter History and Growth | Wikipedia
4. The Streaming APIs | Overview
5. Recursive Deep Models for Semantic Compositionality Over Sentiment Treebank | EMNLP
6. Understanding Sentiment Analysis: What It Is & Why It’s Used | brandwatch

THANK YOU