Analyzing Real-Time Streaming Twitter Data

Syed Zain Raza, Hadia Hameed, Sushmith Ramesh
Instructor: David Belanger

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Motivation

- Businesses can use tweets to gain insights into customer needs. These insights can be used for understanding "Brand Perception" to improve products and launch new services.
- Organizations can connect with their target audience on a more personal level by leveraging their understanding of human thoughts and experiences through the analysis of social media interactions.

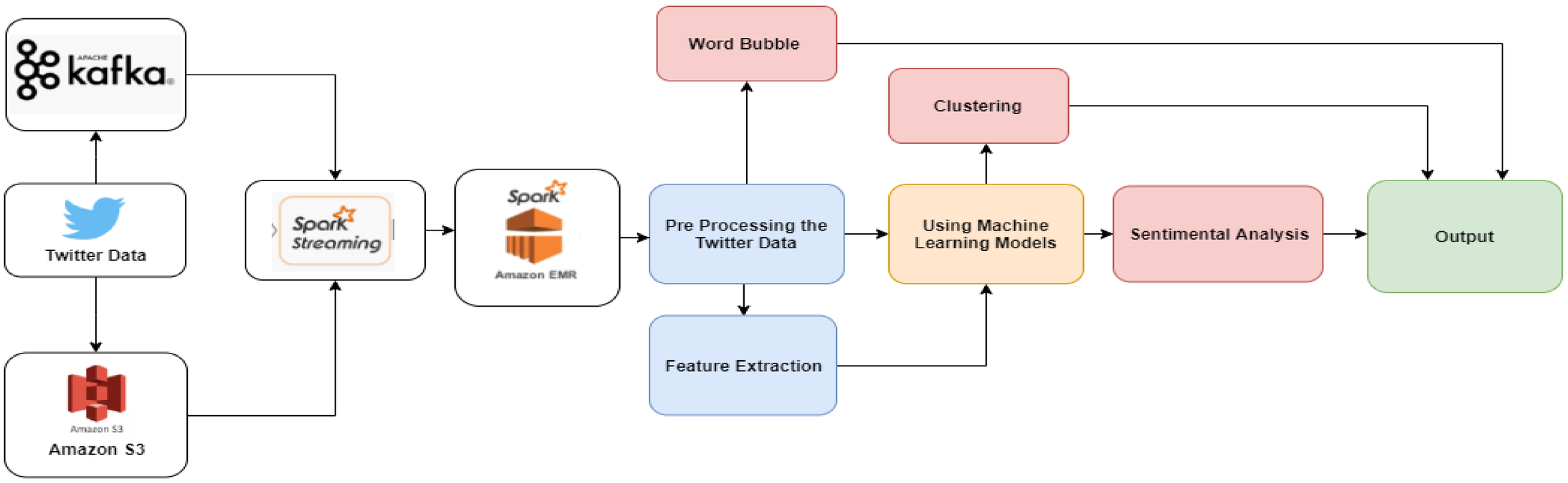
Technology

- **Twitter API:** Get streaming and static data.
- **Spark / Kafka:** Stream & perform data parallelism.
- **Amazon EMR – EC2**: Scale and increase the number of clusters achieving performance gains.
- **Python - NLTK:** Data preprocessing and visualization using Matplotlib.
- **Spark MLlib**: Use machine learning algorithms for sentiment analysis and clustering.

Current & Future Work

- Real time sentiment analysis to understand given trends.
- Visualizing word bubbles showing most frequent words for a given hashtag.
- Clustering tweets to compare user behavior and responses.

Architecture



Methods / Algorithms

- TF-IDF algorithm was applied to convert textual data to feature vectors
- K-means clustering using Euclidean Distance
- Removing stop words and punctuation from the streaming tweets and selecting top N words to make a word frequency bubble.
- Naive Bayes, Decision Trees, Logistic Regression and Deep Learning models for Sentimental Analysis
- Dividing the data into chunks to compare training time for different algorithms.
- Comparing time taken to train and perform analysis on single and multiple instances