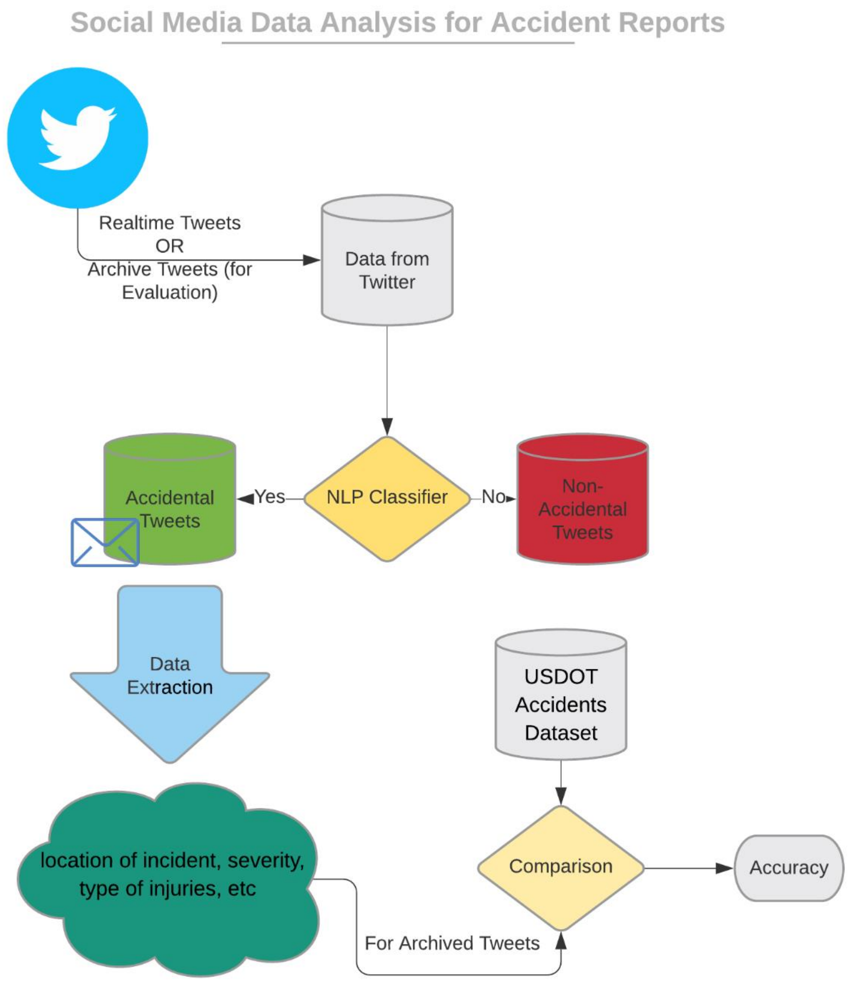
Objective: Social Media Data analysis for Accident Reports

Description: In this project we will collect social media data (from Twitter) and analyze it to retrieve any roadside traffic accident related information (like the location of incident, severity, type of injuries, etc). A particular region will be specified to focus on for data collection.

Team: Abhishek Jajal (APJ180001), Abhisek Banerjee (AXB180050), Divya Sharma (DXS180031), Matthew Koch (MJK180004), Zheheng Zhao (ZXZ163930)

Programming environment: Python, Twitter API, Spark streaming, Kafka, NLTK.

Project Pipeline:



Tentative Method:

1. Data Gathering:

To collect historical data and real-time data from Twitter. Historical data will be used initially in order to refine and create the model. When the algorithm can sufficiently process a test historical data stream, we will stream real-time data. Tweets from public Twitter accounts owned by public agencies and media groups and. individual users would contribute to the data we would require. We are of the opinion that majority of the tweets regarding accidents will be posted by public Twitter accounts rather than the individual users who will have comparatively less contribution in this regard.

2. NLP classifier:

NLP classifier would classify a specific tweet into either an accident tweet or non-accident tweet. This phase would involve understanding and analyzing different Natural Language Processing classification algorithms and identifying the technique that suits our model and gives us desired result. A dictionary of relevant keywords and their combination would be helpful in classifying the tweets.

3. Data Extraction:

To extract the details regarding the accidents like the location of incident, severity, type of injuries, etc. from the relevant tweets.

4. Project Evaluation:

In order to measure the accuracy of our implementation we are going to cross check the accidents with the datasets provide by U.S. Department of Transportation (USDOT)/Bureau of Transportation Statistics' (BTS') National Transportation Atlas Database (NTAD), i.e. https://data-usdot.opendata.arcgis.com/datasets/fatality-analysis-reporting-system?geometry=-97.685%2C32.576%2C-95.906%2C32.980

Future Scope:

Enhancing the project to not only use text but images also as the data for detecting Accidents. It would require additional technologies to identify images that are related to accidental incidents and extract details about the accident from the visual data source.

Tentative Schedule:

Week 1 & Week 2	Data Gathering
Week 3 & Week 4	NLP classifier and Data extraction
Week 5	Project Evaluation
Week 6	Final submission and presentation

Related Literatures:

- Finding and Tracking Local Twitter Users for News Detection by Hong Wei, Jagan Sankaranarayanan, Hanan Samet - http://www.cs.umd.edu/~hjs/pubs/sigspatial2017-localnews.pdf
- TEDAS: A Twitter-based Event Detection and Analysis System by Rui Li, Kin Hou Lei, Ravi Khadiwala, Kevin Chen-Chuan Chang - https://ieeexplore.ieee.org/document/6228186
- Real-Time Traffic Incident Detection with Classification Methods by Linchao Li, Jian Zhang, Yuan Zheng, Bin Ran - https://www.researchgate.net/publication/318406918_Real-Time_Traffic_Incident_Detection_with_Classification_Methods