1. **Architecture**

The proposed architecture for analyzing the social has the following components:

1. Social media data crawler

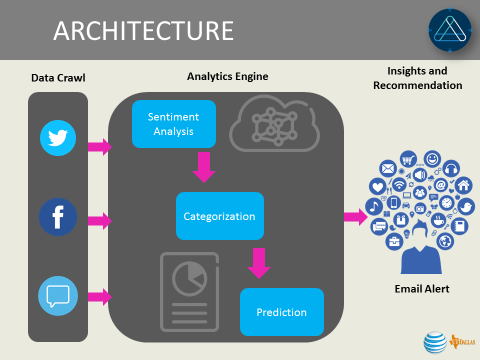
* It connects to the social media sites through *oAuth*
* Fetches the data from the sites periodically

1. Analytic Engine

* Train the tweets using the machine learning algorithms
* Predict the sentiments of the live tweets

1. Insight provider

* Based on the results from the prediction, the insight provider takes decisions to alert the ATT management based on the severity.
* It also identifies the stores based on the zip code and alerts them.

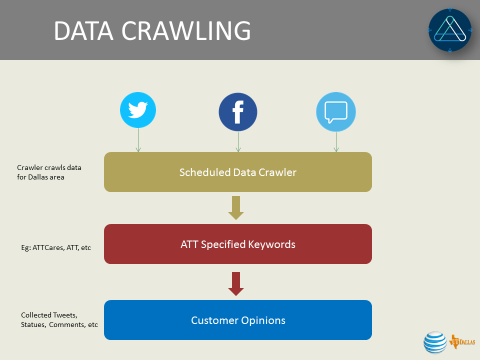


1. **Steps involved in fetching the twitter data**

Data crawling from Twitter:

1. Create a sample twitter account
2. Create the following in the apps.twitter.com

* Consumer Key
* Consumer Token
* Access Key
* Access Token



* 1. Create oAuth to twitter using the above credentials
  2. Get a list of zip codes along with their latitude and longitude intersection in the Dallas area using “zipcode” package in CRAN library.
  3. Search the twitter for tweets using the search string “ATTCares” for a particular zip by specifying a radius around that zip.

1. **Sentiment Analysis**

**Unsupervised:**

**Data cleaning:**

* Eliminate URLs from the tweets
* Remove emails from the tweets
* Remove twitter tags
* Remove twitter user names
* Remove white spaces

**Predict Sentiments**

* Train the **polarity** algorithm using the past tweets.
* Predict the Sentiments using polarity text mining algorithm for the live tweets.

**Supervised:**

**Sentiment classification**

Manually classify sentiments as positive, negative, and neutral for a training set of tweets.

Clean the data [removal of special characters, lower cases, stop words and stemming]

**Model training**

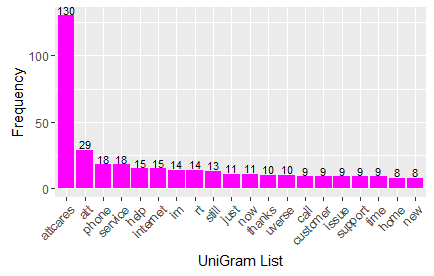
Then trained model using “Random Forest” and “XGBOOST” and then predict the sentiment score for the test tweets.

1. **Rank zipcodes:**

Aggregate the results from the above learning methods and rank the zip codes based on the negative tweets.

1. **Results from Sentiment Analysis**

**Unigram analysis**



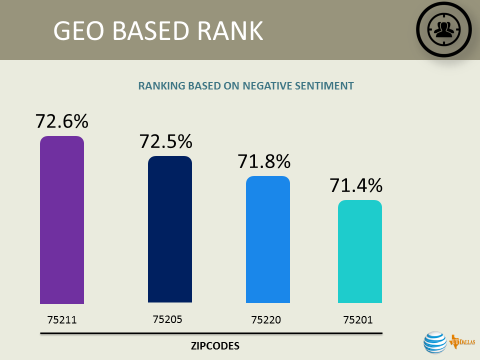
Above represents unigram for 158 tweets collected in the Dallas region. As expected, ‘attcares’ and ‘att’ tops the list. But, they don’t say much about the nature of the issue or sentiment of the customer. Next, we can see that phone has issues. After going through the tweets, it is the ‘iphone’ users who are actually complaining. Then comes ‘service’ which tells us that people are complaining about the ATT service and seeking ‘help’. Then comes the ‘internet’, which tells that the internet users tops the list after iPhone users. Then comes the ‘still’, which means that the customers are facing the problem even after reporting. Then comes the ‘uverse’, which tells us that the U-verse users are facing the problem.

Ranking the complaints based on the unigram:

* Iphone
* Internet
* U-verse

**Ranking of zip codes**

The output from the prediction algorithm is ranked based on the negative sentiments on a zip code basis. The below chart shows the rankings for the top five zip codes with the highest negative sentiment being given first rank.



**Proposed work:**

Consider the negative tweets from each zip code and categorize them as follows:

* LOW SEVERITY
* MODERATE SEVERITY
* HIGH SEVERITY