1. **Abstract**

We used Twitter feed data to be able to make an assessment of the University at Buffalo’s online presence and brand image with the help of sentimental analysis. Universities are large organizations. And just like a multinational corporations image and reputation impacts its sales, a University’s image and reputation plays a vital role in the kind of students, staff and employers it attracts. Until recently, the PR cells of large organizations employed a few interns to handle their online presence. But as their business grew, they adopted sentiment analysis techniques to do this, reducing the efforts on the part of their staff. We would like our University to ride the wave and integrate sentiment analysis techniques into their PR machinery and in the process be one of the few educational organizations to do so, early on.

Results:

UB has a positive image in the eyes of the “Tweeple”

Real life events do affect the positive/negative sentiment and even though the positive sentiment is higher, it is not that far above the negative sentiment

1. **Introduction**

As per Pew Internet Project’s study, about 74% of adults use social networking sites. Twitter is a popular platform used by people to express themselves without inhibition. Tweets are mostly, unbiased opinions on various topics, current issues, their criticism, product reviews etc. Therefore, the microblogging website is a reliable source of true sentiment that a brand, organization or event evokes in people.

This semester an incident allegedly involving racism came to light that agitated UB students belonging to a particular community. The news broke out on a social media and many outraged students took to the microblogging website to vent their anger. This would have created a lot of negative publicity for UB, and its reputation would have been tarnished in the eyes of the users who follow these students and read their tweets; many of whom would be planning to attend UB in the future. Had the university been equipped with a system that could monitor these peaks and troughs in relevant Twitter activity, it could have responded swiftly and mitigated the situation. Our objective was to build a system that would answer these problems or at least be a step in the right direction towards achieving this goal.

* 1. **Contributions:**

Pranay worked on configuring Flume, the extraction of twitter feed via Twitter API and the subsequent ingestion into HDFS, loading the data from its unstructured form to a structured, but raw form into HIVE, cleaning the data and loading relevant columns (id, date of creation, text, screen name), computing the sentiment for the tweets.

Sourav worked on loading the data dictionary into a HIVE table, mapping the sentiments to the original tweets, moving data into excel and then creating PowerView reports based on the data stored in this excel.

**Related Work / Novelty of your technique**

Sentimental analysis from Twitter data is a topic on which some extensive research is done and there are various published works for the same. For the purpose of the project we have referred to “*Lexicon-Based Methods for Sentiment Analysis paper by Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll and Manfred Stede” and* “*Semantic Orientation CALculator (SO-CAL) uses dictionaries of words annotated with their semantic orientation (polarity and strength)”.* The paper talks about the various methods of performing Lexicon based analysis. The paper also talks about the different datasets and dictionaries available. Our project is a re-implementation of the simple lexicon based approach used for sentiment analysis. Although it is not as accurate as the machine learning approach wherein an algorithm is trained to classify data, it is the preferred approach for handling big datasets as training an algorithm takes a lot of time.

1. **Design Issues**

Our design relies on a predefined polarity assigned to words against which the words in a tweet are matched to calculate the sentiment of that tweet. This approach has the following shortfalls:

* Calculates the sentiment based on words. This makes it inefficient while handling negation. For example, the following tweet “University at Buffalo. Not great or even good”. Sentiment for this tweet would be computed to be 2 (1 for great plus 1 for good), which is a positive score even though the underlying sentiment of the tweet is clearly negative.
* The dictionary contains 8221 words, which is not exhaustive and does not take into consideration phrases and contextual meaning.
* Underlying assumption that tweets would contain grammatically correct language. However, a limit of 140 characters makes sentiment analysis for microblogs challenging because of problems like use of short hand status message, informal words, word shortening and spelling variation.

1. **Algorithm Description**

To assign a sentiment score to the tweets we have used the lexicon-based approach. Lexicon Based techniques work on an assumption that the collective polarity of a sentence is the sum of polarities of the individual words. For implementing this approach, a dictionary needs to be used. A dictionary is a list of words classified as positive, negative or neutral. We obtained this from the web. Our dictionary has a list of 8221 words consisting of nouns, verbs, adjectives etc.

* We first converted the Twitter data from its raw form into a tabular format.
* The text part of the Tweet stored in the table is then split into its individual component words.
* The words are looked up in the dictionary and the corresponding polarity is fetched and stored in a second table.
* If the sum of the polarities of these words is positive, then a “positive” sentiment value is assigned to the tweet. If the sum of polarities of these words is negative, then a “negative” sentiment value assigned to the tweet. Else the tweet is assigned a neutral sentiment value.
* These sentiment values are stored in a new table with tweets.

Due to infrastructure constraints we have implemented a single node Hadoop installation. However, the distributed nature of our project is in the storage of the tweets extracted. The idea here is that this approach could be scaled up in the following two methods:

* The tweets could be extracted on multiple compute systems brought together in one system and processed using lexicon based approach.
* Alternatively, twitter data could be extracted on multiple compute systems. The tasks involved in computing the sentiment of the tweets could be distributed among the individual compute systems.

The lexicon based approach for calculating twitter sentiment takes 13 seconds (considering only the time taken for computation of sentiment) for 1051 tweets and 4 HIVE queries. It is important to note that the time taken could be alluded to the fact that a virtual machine was used. The same approach would perform much better if a local installation is used. The appendix contains the logs for the queries.

1. **Data Sets and Software**

Dataset includes tweets extracted using the Twitter API starting from 25th November 2015 to 7th December 2015. “#UBBulls”, “SUNY Buffalo”, “University at Buffalo” are the keywords used for extracting relevant tweets. The data is in English language only. It contains retweets as well. We used HDP1.3 distribution on Hortonworks Sandbox, a virtual machine based on CentOS. The Hadoop distribution comes preinstalled with Flume and HIVE. The VM is run using Oracle VM VirtualBox Manager.

1. **Empirical Results**

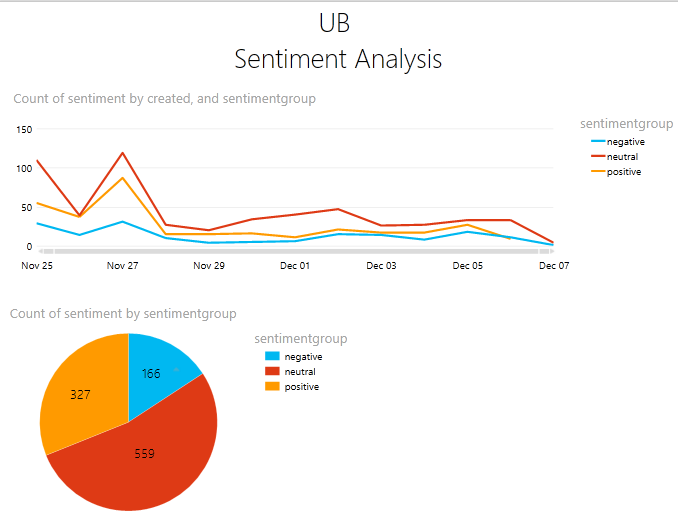
Below is the graph presented in Excel Powerview with the help of processed data exported from HIVE table. It provides a summary of the results of our sentiment computation performed on tweets related to UB.

1st graph – Line graph – Shows the rise and fall in the number of positive, negative and neutral tweets with respect to time.

X – axis: Dates on which the tweets were created.

Y – axis: Count of tweets. Where each line represents a particular sentiment group.

2nd Graph – Pie Chart – Shows a distribution of the number of tweets of all the three sentiment groups.



The following observations can be made using the above pie chart,

* The number of positive tweets are higher than the number of negative tweets. Thus between November 25th and December 05th UB has had a positive image on twitter.
* There was a spike on 27th November owing to a football game on that day. People supporting the team generally tweet encouraging and motivational things up until the time leading to the game, which explains the rise in the positive tweets. But UB lost that game and therefore there is also a rise in the number of negative tweets.
* There is a high number of neutral tweets. This could be due to the fact that many tweets cannot be classified as either positive or negative as they contain only information in which case the text would not include any positive or negative words.

1. **Practical Experiences**

One of the issues we faced was the infrastructure setup. This consumed considerable amount of time. The HDFS file structure changes drastically between releases of Hortonworks Sandbox. We had initially installed HDP 2.3.2 wherein we faced some issues trying to configure Flume. We were unable to resolve this issue so without wasting anymore time, we decided to deprecate to HDP 1.3 and easily found more material on Flume setup. This did not affect our project in anyway and we were able to achieve our objective using an older version of the Hortonworks distribution. We also learnt that lexicon analysis is not very accurate due to various reasons mentioned earlier in this report. If we had infinite resources, especially time, we would have like to explore a more efficient approach to lexicon analysis, one that handles emoticons, negation and blind negation. Emoticons are generally used by people around the world to depict emotions. They are language agnostic. Hence it is obvious that they carry very useful sentiment information in them.

1. **Conclusion and Future Work**

Results we obtained from our project as mentioned below:

UB has a positive image in the eyes of the “Tweeple”

Real life events do affect the positive/negative sentiment and even though the positive sentiment is higher, it is not that far above the negative sentiment

The accuracy of the lexicon based approach depends, to a large extent, on the dictionary being used and the way it is being used. Our dictionary did not include plain text emoticons.

We really enjoyed working on this project as it was our first exposure to unstructured data and Hadoop. With respect to sentiment analysis, we would really be interested in extending our work to emoticons. We believe they are a universal way of expressing sentiment and is independent of language. They do not involve sarcasm and negation and would help gauge the sentiment more accurately. We would also like to implement sentiment analysis using the algorithmic/ machine learning techniques and do a comparison with the lexicon based approach in terms of accuracy, implementation time and the ability to handle really large datasets. A couple of features we would want to add to the system is to find out the popular words being used in tweets related to UB and identifying social media advocates( people whose opinions about the university have an impact)

1. **Bibliography**
2. <https://www.aclweb.org/anthology/J/J11/J11-2001.pdf>
3. <http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/>
4. <https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>
5. <https://cwiki.apache.org/confluence/display/Hive/StatisticsAndDataMining>
6. <http://blog.datumbox.com/10-tips-for-sentiment-analysis-projects/>
7. <http://blog.datumbox.com/10-tips-for-sentiment-analysis-projects/>
8. **Appendix**

This section provides the snapshots the steps involved in implementing sentimental analysis using the lexicon based approach.

In the embedded files below, are the snapshots of the steps and the Hive queries used.