# NSA – News Summarizer & Analyser

**Nidhi Panpalia** **Rushabh M. Shah** **Siddharth Shah**

Under the guidance of

**Prof. Niranjan Balasubramanian**

Stony Brook University, Stony Brook

**Abstract:**

The speed by which the information is growing on the Internet today has no bounds. For a user to read about particular topic can be painful given the amount of information. When it comes to news for a particular person/topic, almost all news articles on Internet give repetitive information other than few little new details in every article. So we decided to provide a summarization tool which aggregates important information from various news articles available on the Internet and provide a brief summary on it. Also, we further found tweets mentioning this news articles and performed sentimental analysis on them giving the user an overall view of the opinions of public for that news article.

.

**Introduction:**

Today the world is shifting from paper news to electronic news articles. With people's busy schedule, it is difficult to go through all these news articles and also many of the news websites are majorly focusing on advertisements for commercial purposes rather than giving succinct relevant information about the topic. To remove redundant data and getting the important gist of the topic, a news summarizer which will aggregate important information from multiple websites is a perfect solution. The proliferation of social media in the recent past has provided end users a powerful platform to voice their opinions. Businesses (or similar entities) need to identify the polarity of these opinions in order to understand user orientation and thereby make smarter decisions. This gave us the motivation to further provide a sentimental analysis to the user on his/her search query.

*Figure 1: The Internet - as a major source of news*

**Features:**

1. Search engine UI
2. Clustering of related articles
3. Summary of relevant news articles
4. Manual labelling of Twitter Hashtags
5. Sentiment Analysis of tweets mentioning the news articles
6. Graph plot on the positive tweets vs negative tweets
7. User friendly, clean simple to use interface

**Design:**

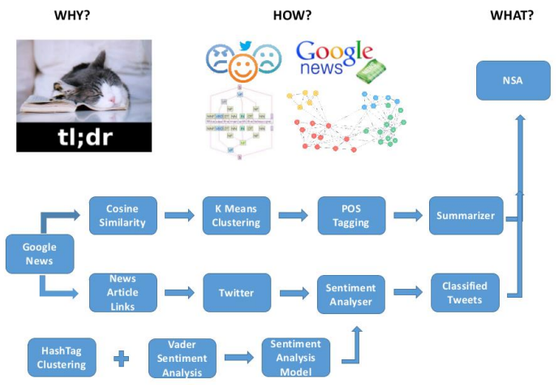


*Figure 2: The general process flow diagram*

We first take the search query from the user, this search query is passed to Google News and urllib2 library to open the links. Then we download the HTML page and convert to plain text format. We have set a limit on the number of articles which will be fetched from Google News. We then use Cosine similarity to find documents which are relevant to each other. These documents are then clustered using k-means algorithm. For summarization, we use POS-tagging, TF_IDF and tex-rank algorithms have been combined and used. We then pass the URLs obtained initially to Twitter API which starts the search for the tweets which have mentioned these news articles. These tweets are then analyzed for positive and negative sentiments. To train the sentiment analyzer, we created a labeled dataset from streamed twitter data for few search terms. For the labeling we used 2 techniques: Manual labeling and Vader's Labeling technique. Model trained using manual labeling helped us correctly label the tweets sentiment for the person. This reduces the error of classifying the tweet as positive or negative for the person just because he/she is mentioned in the tweet.

Later we have used Vader's Labeling for labeling the general search terms. For classifying, after trying various classifiers, we found the SVM-linear classifier gave the best result.

**Result:**

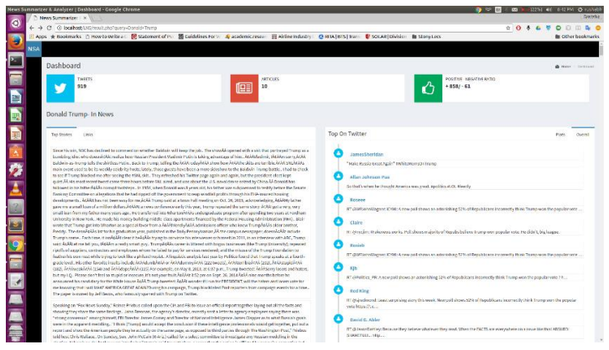


*Figure 3: The developed application*

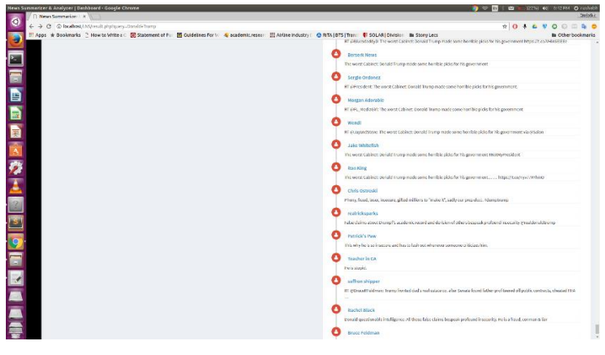

*Figure 4: Tweets analysed to be negative*

**Conclusion:**

In this application we have tried various techniques to improve the summarization of news articles fetched from Google news. We used clustering algorithms to cluster news articles which are similar before summarizing them. Also, in sentimental analysis we have tried to introduce a new way to train the model to label tweets using manually labeled tweets. This improved the classification of tweets to negative and positive tweets. After trying various classification algorithms we found Linear SVM worked the best.

**Future Scope**

As a future scope for this application, summarization can be improved by abstractive summary rather than extractive. Also, currently the dataset we are getting from twitter is quite redundant because it is retweeted many times and it is fetched again. As a solution to this we can apply similarity algorithm on tweets and if similarity is more than 0.95, then we can ignore those tweets. Another possible work can be that the process of manually labelling can be done for a wider range of generic news topics.

**References:**

1. "A survey of Text summarization Techniques" "https://www.cs.bgu.ac.il/~elhadad/nlp16/nenkova-mckeown.pdf
2. Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).
3. Vader https://github.com/cjhutto/vaderSentiment
4. POSTagger http://thetokenizer.com/2013/05/09/efficient-way-to-extract-the-main-topics-of-a-sentence/