STOR 320 Project Proposal

Troy Hall (Captain), Rees Braam (Doctor), Sam Galloway (Engineer), Sidh Kulgod (Emissary)

Proposed Questions:

1. Finding out if the complexity of a news story (unique words, word length) has anything to do with the orientation (positive/negative) of the article.

Identifying biased news sources as an extension of this question is also a possibility, using keywords to determine what articles are about and if they are positive or negative. We are going to construct a list of “positive” and “negative” words and then match them to the text of each articles in order to determine the overall sentiment of each article. We also understand that words like “bad” and “negative” have much lighter meaning than words like “disastrous” or “cataclysmic,” so we are also going to utilize categories of words when creating the lists to properly weigh the attitude of published content. The chosen data sets are made up of many news stories, and we feel that we have the required abilities to crawl the links supplied by the dataset in order to obtain text to properly analyze.

1. Does the date or day of the week an article is posted have an effect on its length/complexity?

We wanted to see if the date/day of the week that an article was posted influenced its complexity; for example, to find if news organizations post stories with lower word counts on weekends or days that they feel that more people will be reading. Our hypothesis for this question is that more digestible news stories (shorter length, fewer unique words) are posted on Friday/the weekend to capture greater interest.

1. Compare number of “positive” and “negative” articles posted by month, by organization to determine if overall news organization sentiment went up or down over the sampled period.

By checking which organizations became more, or potentially less negative over time, we hope to determine which news organizations function for criticism and which, if any, exist in more of a state of celebration.

1. What is the optimal balance of multimedia to text to ensure maximum popularity (measured in number of shares or association)?

By using the links from one of the datasets and the given data of shares per article, we can determine the best ratio of multimedia (photos, videos, and embedded graphics) to number of words in each article and use the resulting ratios to find if there is a significant effect on the number of shares received in turn.

1. Which news organizations publish the most “positive” stories and which ones publish the most negative ones?

This is the initial step for one of our previous questions. By using the sentiment score that we will create and refine in each dataset, we can determine which organizations are more positive or negative overall over the sampled period. While determining how they change over time will be an added level of interest, we also expect determining net sentiment for each organization to allow us to cluster organizations into a “positive” group and a “negative” group.

Data/Data Sources:

1. <https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity>
   1. The UCI Machine Learning Repository retrieved this dataset back in January 2015. It gives many statistics of Mashable articles from the retrieval period. We are confident that we can add the text to the data frame using the links provided in the leftmost column of the data. We also feel like Mashable is a good representation of social – oriented news as opposed to hard news.
2. <https://www.kaggle.com/snapcrack/all-the-news>
   1. This dataset was compiled by a Kaggle user named Andrew Thompson from 2016 and 2017, with a number of articles coming from 2015. With no links to each article but the raw text, title and time data available, this dataset gives the raw data that we need to make sentiment scores and to work with the content of the articles. This will be used for the questions that just require the text to supplement the first and third datasets, and to test any predictions that we make about the content of certain news organizations.
3. <https://www.kaggle.com/uciml/news-aggregator-dataset>
   1. This final dataset was also retrieved by UCI Machine Learning and has an abundance of other organizations than the Mashable dataset. It also crucially includes the URLs so we can yank more information from the articles as we need it with requisite code. This also has very close dates with the other datasets, with the information being retrieved in mid-2014, increasing its integrity.

Methods and Scope:

1. We are thinking about crawling the Online News Popularity dataset links to add text to the observations, which would make analyzing the data much easier.
2. Adding a sentiment score to the “all the news” dataset is required for some of the questions that we want to answer, Rees has already done some preliminary analysis and it is viable to add the score.
3. Most data that we need is accessible through the html files of the webpages, and we intend to use them to a.) verify the accuracy of the data and b.) create our own columns when we need them. This will be possible since we picked multiple datasets that don’t just have the text, but also have a route to obtain more information from the stories.
4. We would like to find some social media data from the same time period and compare it to the overall sentiment of media to see if there was discord between the attitudes of media organizations and the things that people were posting on a more personal basis. We’ve been able to find ludicrously large datasets that had massive amounts of data from individual websites in them (ex. Over a terabyte of data that holds all of the comments from reddit from 2007-2015), but we have not been able to find anything that is of a manageable size. We will continue the search in the initial of the stages of the project to find a suitable dataset, but if we fail in the search, it is not critical to any of the questions that we have postulated in the proposal.

Brief Description of Related Work

Media insight companies have become a huge part of the global economy, helping companies to aggregate the glut of data that may be produced about their brand on a regular basis. Companies like Brand24 and Media Sentiment Inc. are already utilizing/monetizing this sort of technology to add value to their customers. We believe that we can use this data almost as a watchdog to make insights about media organizations that are hopefully accurate and useful in determining potential biases that may stem from the modern digital popularity contest.