

## BUDT758T: Data Mining & Predictive Analytics

### Data Mining for Business (BUDT758T)

Project Title: **ANALYZING THE MEMETRACKER DATA**

Team Members:

Sreerag Mahadevan Cheeroth

Allika Thadishetty


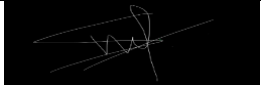
Sneha Chandrashekar

Pakshal Shah

Arvind Kanhirampara Ravi

### ***ORIGINAL WORK STATEMENT***

We the undersigned certify that the actual composition of this proposal was done by us  
and is original work.

| Contact<br>Author | Typed Name          | Signature   |
|-------------------|---------------------|---|
|                   | Sreerag M Cheeroth  | Sreerag M. Cheeroth   |
|                   | Allika Thadishetty  |  |
|                   | Sneha Chandrashekar |   |
|                   | Pakshal Shah        | Pakshal Shah  |
|                   | Arvind K Ravi       | Arvind. K.  |

## II. Executive Summary

In an endeavor to thoroughly analyze and decipher the voluminous memetracker dataset, encompassing around 96 million records with a monthly data range of 1.5-2.7GB, we undertook a meticulous study of several months' data. Our method involved extracting and interpreting "Quotes" using a text mining process.

The pivotal focus of our investigation was to comprehend the process of labeling the sources of information based on the polarity of sentiment conveyed. By monitoring these sources over diverse periods, we aimed to track their evolution over time. This strategy provides a deep understanding of an information source's performance on the internet. With some modifications, this concept can be harnessed to identify sources giving strongly negative sentiments, signaling domains that may be best to avoid.

To further amplify our understanding, we attempted to turn this information into a visually engaging network diagram. This approach would allow us to discern how positive and negative domains interact, made possible by the inclusion of 'L' in our dataset, representing hyperlinks referred to in posts. By identifying these interconnections, we could effectively categorize them, thereby enhancing our comprehension of the data's intricacies. This study provides a robust foundation for further explorations in this domain.

## III. Data Description (1 page)

Professors and students from Stanford had built a large scale framework to track and extract memes through online text. They had developed scalable models and analyzed 1.6M mainstream sites, 90 million articles for a period of 3 months.

We used this data as our source for the project. The dataset covers a period of data from August 2008 to April 2009. Our focus was primarily on data from August 2008, October 2008, and December 2008, because these files were smaller than 1.25 GB each.

original dataset: <https://snap.stanford.edu/data/memetracker9.html>

## III. Research Questions (1 page)

Describe the questions that you plan to investigate using your data. Use the terminology of the business area that is relevant (not statistical hypotheses). These should include questions that are related to classification, prediction, etc.

## IV. Methodology

### Data Cleaning and Pre-processing

We used Python as our programming language as our data set was huge and it wasn't possible to work with such a large data on R (data sets; quotes\_2008-08.txt.gz, quotes\_2008-10.txt.gz, quotes\_2008-12.txt.gz). It was a task to clean the data as it was similar to a long pivot.

| 0    | 1   |
|------|---|
| 0 P  | eighteezbaby.com                                  |
| 1 T  | 2009-01-01 00:00:50                               |
| 2 P  | thelondonreviewer.com                             |
| 3 T  | 2009-01-01 00:01:10                               |
| 4 Q  | not only do they patients get to lose their lo... |
| 5 L  | entertainment.timesonline.co.uk                   |
| 6 P  | thelondonreviewer.com                             |
| 7 T  | 2009-01-01 00:01:10                               |
| 8 L  | imdb.com  |
| 9 L  | imdb.com  |
| 10 L | imdb.com  |
| 11 P | thelondonreviewer.com                             |
| 12 T | 2009-01-01 00:01:10                               |

From the first P to the subsequent P is one data set, we cleaned the data by removing the data sets which did not have a Q value.

## Text pre-processing

In text preprocessing, we processed the text data in Q columns. We have performed text preprocessing on the Q columns using the preprocess\_text function, which converts text to lowercase, removes punctuation and non-alphanumeric characters, removes single characters, removes multiple spaces, tokenizes the text, and removes stopwords. We also cleaned the url to remove the domain name using the parse\_url function. We then cleaned the text to remove the stop words, punctuations, alphanumeric characters, single characters, multiple spaces and also tokenized the text. We then applied this preprocessing function to all the Q's in the dataset using a for loop, which goes through each row of the dataframe and applies the preprocess\_text function to the text in the second column if the value in the first column is Q.

## Sentiment Analysis

After the text preprocessing we made sentiment analysis using the TextBlob python library, the sentiment values for the words were added into a new column in our original data frame and filled the other rows for the column with Null values.

### Visualization and interpretation

Followed by the sentiment analysis we did visualization to draw insights from the results. We plotted Histograms and scatter plot to visualize the distribution of the sentiment values and correlation between the sentiment values and the word count. We observed that most of the words were grouped as neutral as the word count increased more than 25. As we moved over the months the word count decreased overall and the negative sentiment also decreased.

#### IV. **Results and Finding**

##### **Results:**

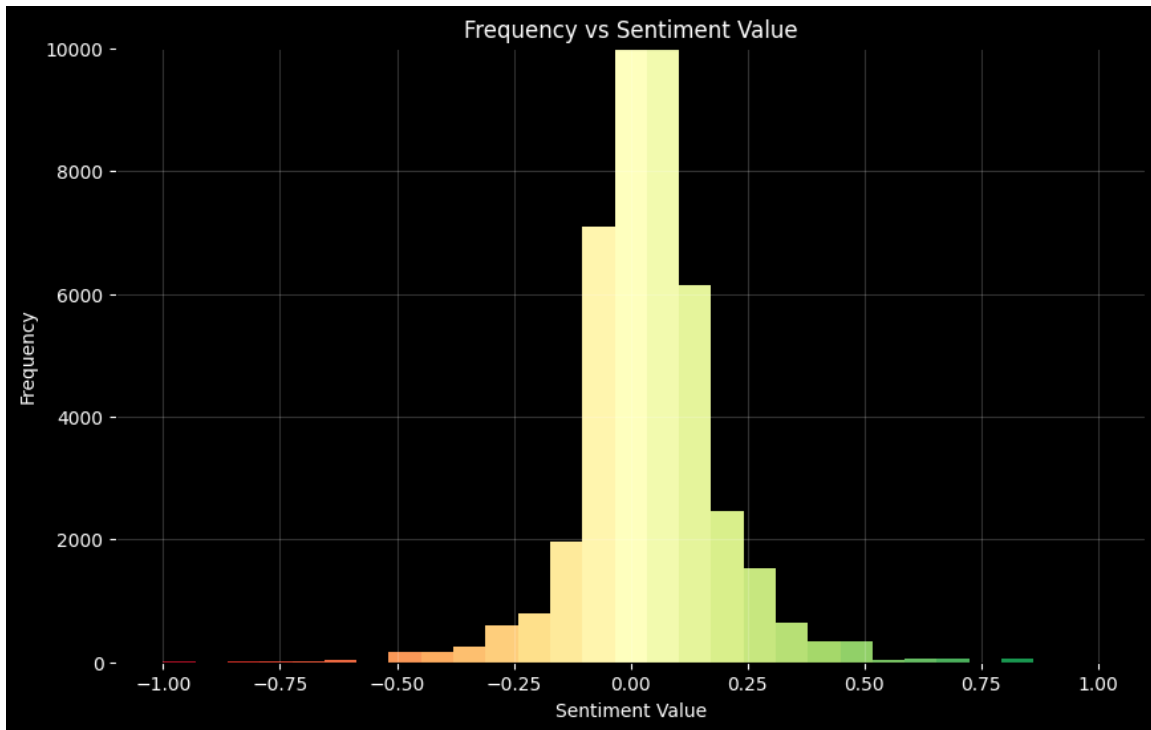
After going through our data in the study, we came across the following results:

Here, due to the computational limits, we had to reduce the scope of our study to just 500,000 rows of the dataset which in itself has over 60 mill+ rows **each**. This huge number of data was hard to handle with just Python or R Studio. In the future scope of this project, we aim to use a database query such as SQLite3 to clean and query the whole dataset to provide higher accuracy in our findings.

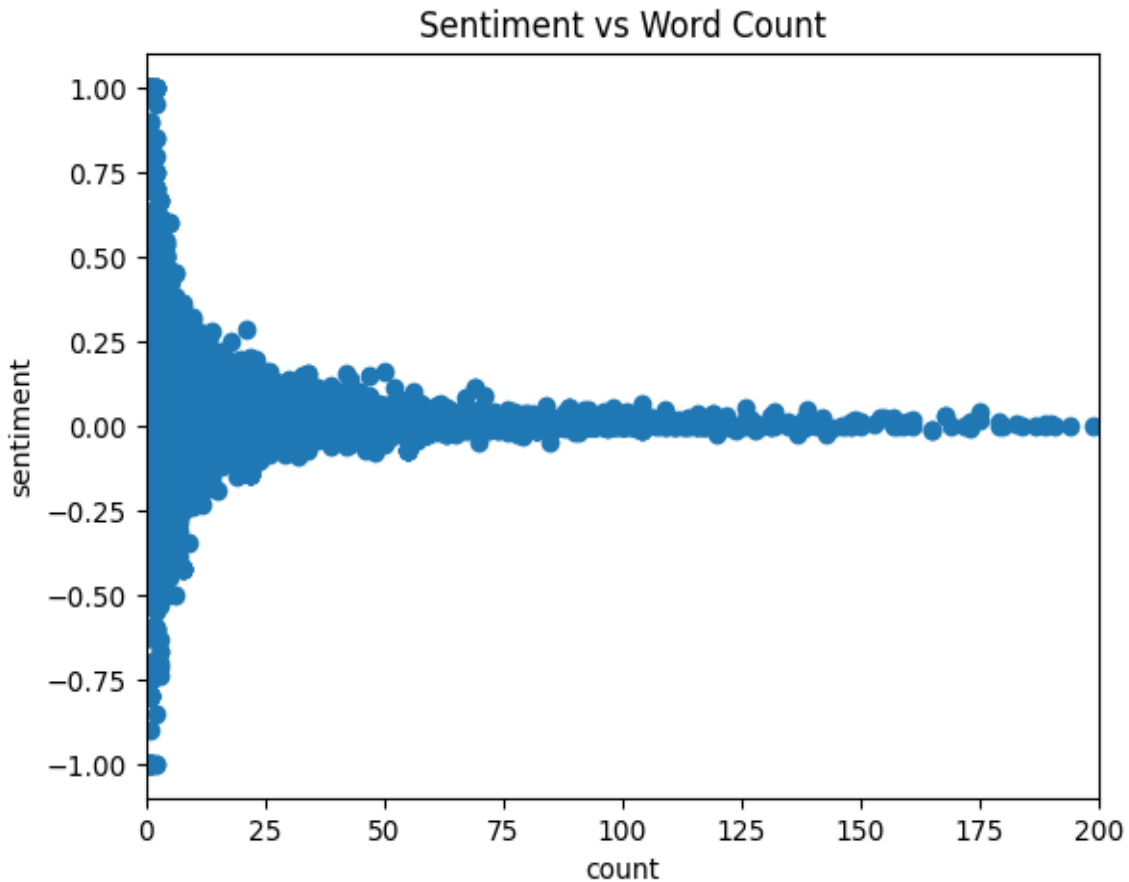
##### **Data analysis for the month of August 2008**

The count of quotes labeled as neutral, positive, and negative neutral is as follows:

|     |       |
|-----|-------|
| =   | 87017 |
| +ve | 47911 |
| -ve | 19754 |



When we look at the distribution of the sentiment values, we can see that most of the data is labeled neutral. When we further compare +ve and -ve sentiment we can see that the negative data takes a bit of upper hand when compared to the positively labeled data that are on the border of change in label. As we move further to the +1 and -1 sentiment scores we can see that the positively labeled data are more in number compared to negatively labeled data. This tells us that there are more extremely positive data than extremely negative data.

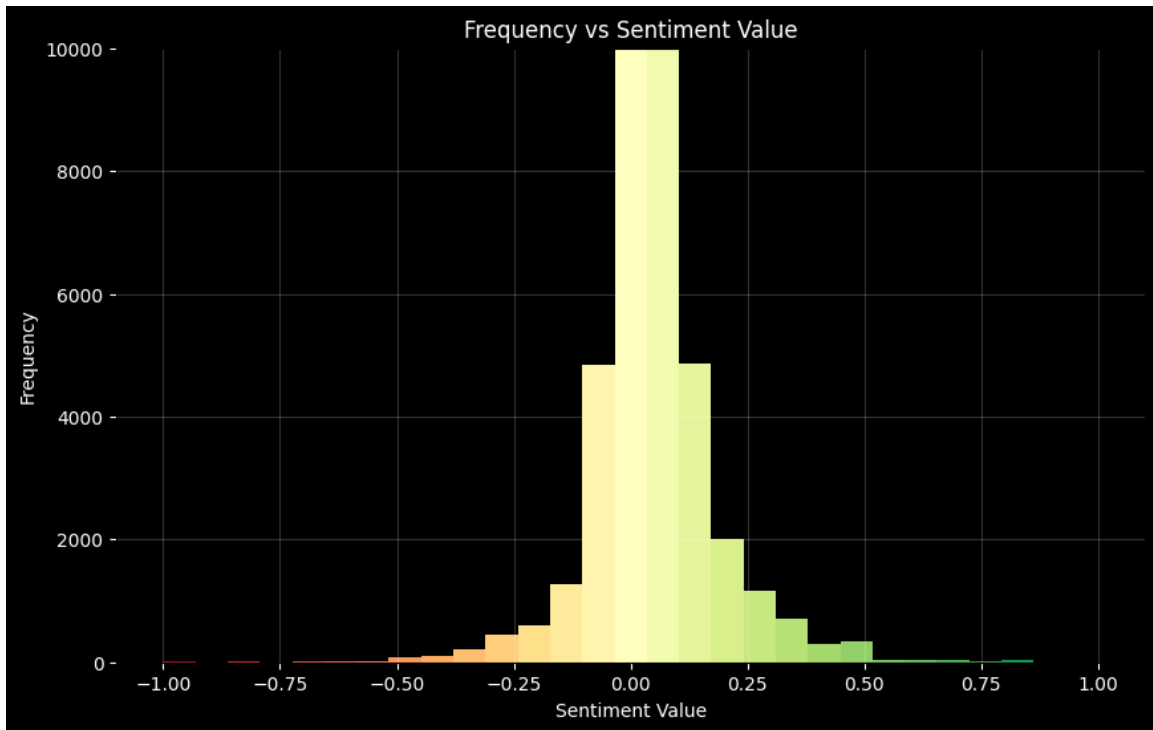


We can see here that as we increase the keyword count from 25 of the text that has been analyzed, they are labeled as neutral. While the text labeled as positive or negative has a word count of less than 25. There might be a case where, when more and more words are processed in a single text the positives and negatives balance out and we arrive at the “neutral” label.

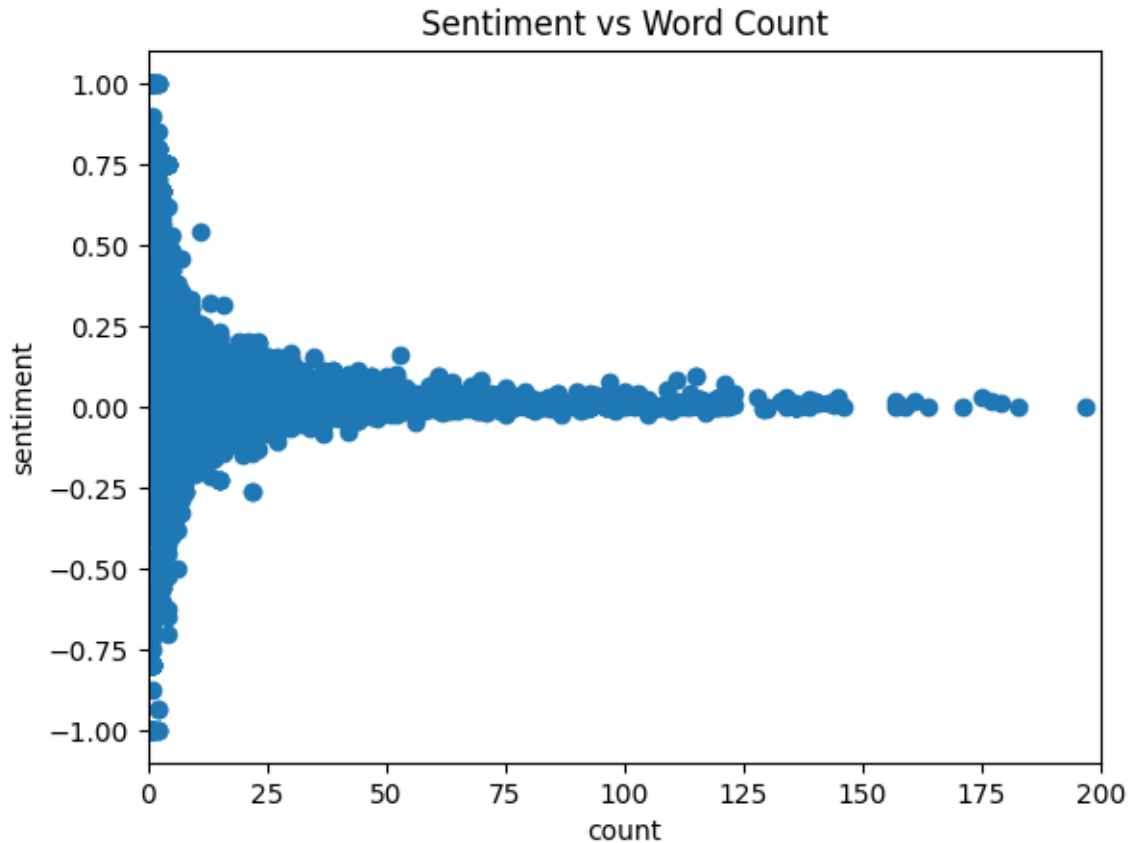
#### Data analysis for the month of October 2008

The count of quotes labeled as neutral, positive, and negative are as follows:

```
= 56633
+ve 36161
-ve 13892
```



When we take a look at month 10, we can see that the neutral data is again predominant but there is a balance between the data that is borderline positive and borderline negative. Looking at the labeled data, between positive and negative data, positive labels are more than negative labels



The overall distribution is the same as the 8th month but here we can see that the text files published this month are a bit less and the neutral labeled data also have a fewer word count compared to the previous month.

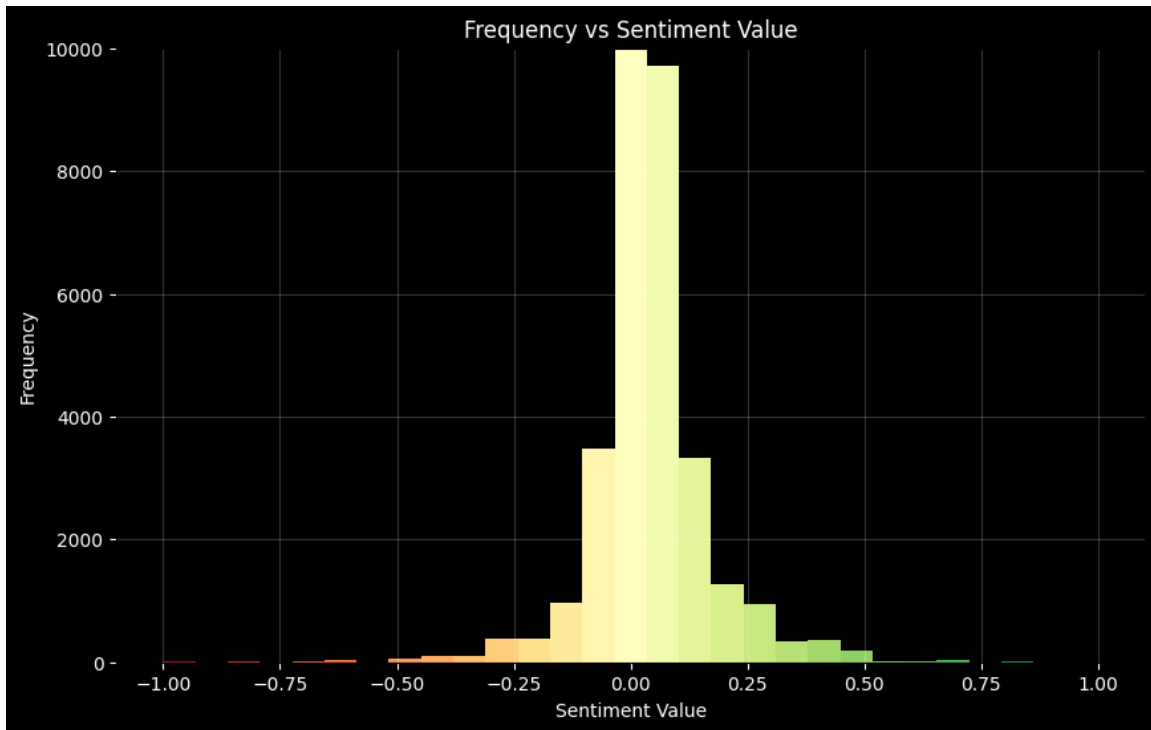
#### Data analysis for the month of December 2008

The count of quotes labeled as positive negative and neutral are as follows:

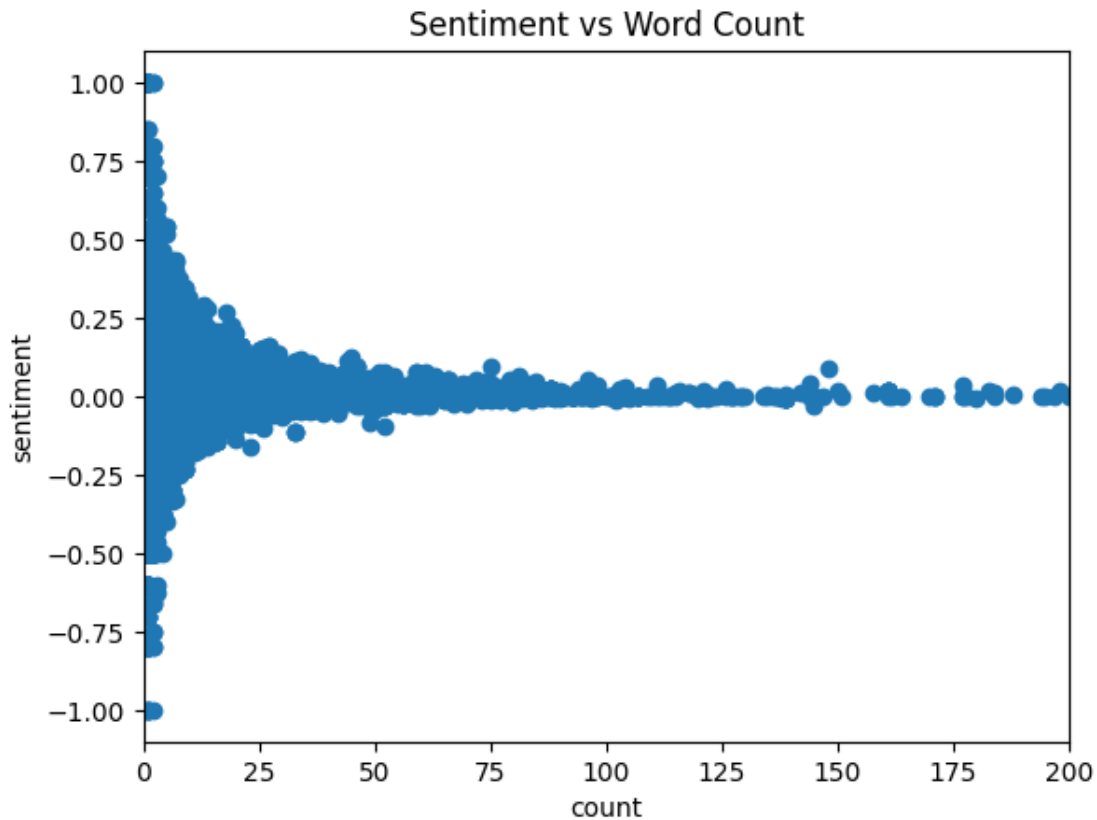
|   |       |     |       |     |      |
|---|-------|-----|-------|-----|------|
| = | 48131 | +ve | 24422 | -ve | 9731 |
|---|-------|-----|-------|-----|------|



## BUDT758T: Data Mining & Predictive Analytics



In month 12, we see that as usual neutral sentiment takes the majority but there is a balance between highly positive and extremely negative data, here also the positive data takes the upper hand just like the last few months.



Here we can see the distribution of word count for more than 25, have data that are labeled as positive and neutral, very few data points are labeled as negative comparatively and most of them have a word count of less than 25. This month there were comparatively less data than in the previous two months.

### Overall Domain comparison

We first tried to take the top 10 positive and top 10 negative sentiment data, but since we are dealing with a subset of the entire dataset we were not able to find common domains within the subset, If we had used the entire dataset we would have definitely found the top 10 domains that share positive sentiment throughout the months and similarly top negative domains.

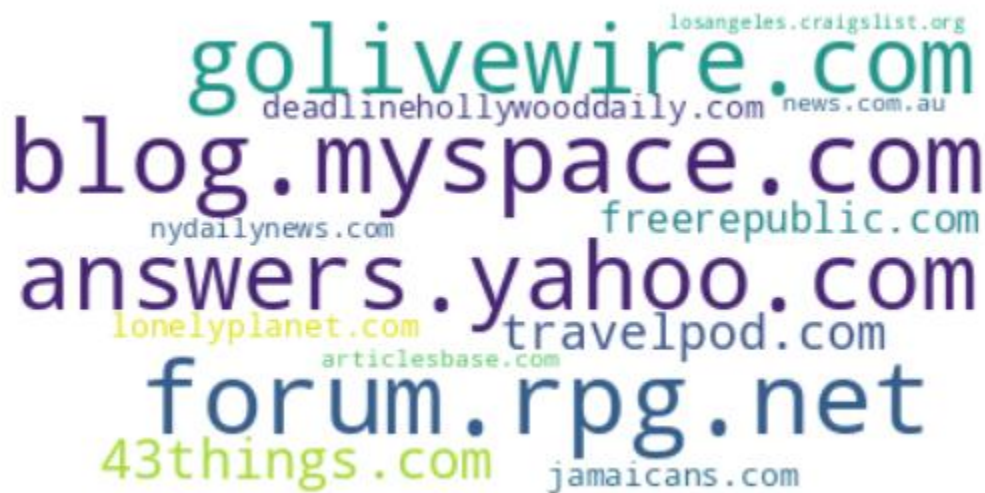
As a reason, we are going ahead with the top 500 most positively scored and the top 500 most negatively scored data from the subset.

We first look at the domain of the top 500 positively labeled quotes



Here we can see that the majority of the positively labeled quotes come from domains like “43things.com” and “blog.myspace.com”

Similarly, we can take a look at the domains that have published the top 500 most negative quotes.



Here, the negative domains are “blog.myspace.com” and “forum.rpg.net”

When we look at it broadly, we can see that some of the domains are common like “blog.myspace.com” and “43things.com”. Here “blog.myspace.com” is a website where people write personal blogs and post them online, since these are personal thoughts and opinions it is natural that there can be negative as well as positive quotes published. Similarly “43things.com” was an online page where people shared their goals, ambitions, and list of things they wanted to do, certainly this resulted in most of the positively labeled data. The

website also had a section to share the hardships and experiences they faced while achieving their goals, this is where the negatively labeled quotes come into the picture.

V. **Conclusion**

We attempted at analysing the sentiment of the quotes published online over the 2008-2009 period. The extremely huge dataset was difficult to interpret and process given the software capacity hindrances.

We decided to move ahead with the subset of data related to 6 months (Aug to Dec) 2008.

We were successful in categorizing the domains based on the sentiment of the content they publish, and also computing the sentiment scores of the quotes published online.

VI. **Appendix (Any additional information to be submitted):**

**Grading Notes:**

The principal criterion is technical quality of the work. In addition, I will grade

- The interestingness and originality of the project. [There is an element of subjectivity here – however this is a criterion on which recruiters and others will judge you and, moreover, the onus is on you to convince me that the work is interesting and original.]
- The ambitiousness of the project. [All other things being equal, choosing a more challenging project will result in higher scores.]
- The effectiveness of presentation. This includes (1) the quality of your PowerPoint deck, and (2) the quality of oral presentation. [Was there a logical flow to the presentation? Did you engage the audience?]
- The quality of writing, and written and visual presentation of statistical analyses. [Clear, crisp and concise writing is rewarded. The report should be professional looking.]
- A portion of the grade will reflect peer evaluations (I will preserve confidentiality of feedback). [Project scores tend to be correlated with team cohesiveness and an ethos in which everyone is committed to making significant contributions.]

**Peer evaluations are to be submitted on Canvas no more than two days after the presentation in class.**