Predicting reddit post popularity

Prasanna Chandramouli
CS, Tandon school of engineering
New York, USA
prasanna.chandramouli@nyu.edu

Radhakrishnan Moni
CS, Tandon school of engineering
New York, USA
radhakrishnanmoni@nyu.edu

Varun Elango
CS, Tandon school of engineering
New York, USA
varunelango@nyu.edu

***Abstract*— Use of socially generated big data, to identify current trends and to predict future trends has become a new paradigm in the field of computational science. A natural application of this would be predictive marketing, which is the process of using data mining and statistical analysis to forecast trends and to implement solution which benefit businesses. In this paper, we endeavor to build a predictive model for the media and news agency whose revenue depends on the contents they publish. We analyze one such discussion forum to identify patterns and trends that would be benefit online marketing. We aim to show that the future popularity of a post can be predicted given the content of the post.**

Keywords—analytics

I. INTRODUCTION

Reddit is a popular social news website, where users post links, text or images which other users can up vote or down vote. This analytic aims to analyze and predict popularity of reddit content. The data was obtained through reddit's api which is free of cost. The data set contains around 17 features or columns with number of ups(votes) being one of them. Our aim is to predict the factors that influence this feature with high accuracy. To predict this, we need to be able to identify the causality of up votes. While a post's popularity is usually related to the post's content, there are often other factors that determine how successful a post becomes. The focus of this analytic is also to analyze how these factors play a role in predicting how popular a post will become. This included features such as the content of the post and its sentiment, the Subreddit of the post, author of the post, and the time of day the post was created.

II. MOTIVATION

In a digital world saturated with web content in the form of entertainment to internet users or business opportunity for marketing companies, only a few items become popular. For example, for every minute users around the world send more than 300,000 tweets [7], share more than 680,000 pieces of content on Facebook [8], and upload 100 hours of video on YouTube [9]. Companies of today's world invest a lot of money in online marketing to boost their revenues. Hence, identifying the digital content that will become popular becomes a matter of foremost importance.

With ever increasing digital content (text and media) on the internet, media companies need analytics to better understand their readers and audience. The way users react to a post should have a big influence in the way companies publish articles and content. Data about user reactions need to be analyzed so companies can adapt and publish posts based on the interests of their user base. This analytic will aim to predict not only the kinds of posts that will draw more reactions but also identify the kinds of topics that appeal (or don't appeal) to a section of users. The plan is to use tools like NLTK for applying NLP techniques on the datasets and make meaningful correlations from user comments, weight of reactions (number of likes/dislikes), topic of the post and other possible features. Data containing user comments and reactions on Facebook or YouTube would be ideal sources for this project but since neither of them have released such open datasets (besides their APIs), websites like Reddit, Hacker News, Stack Overflow would be good alternatives.

III. RELATED WORK

[1] Segal and Zamoshchin attempted to predict the popularity of the reddit post by analyzing only the metadata of the post and not considering the content and the commentary of the post

[2] Bargdon, McMorran, Sahay and Wong aimed to predict the karma score using SVM with limited feature vectors. They were able to predict up to only 42% accuracy as they were handicapped with only 34,000 posts.

[3] Moreno, Redondo2 authors talk about the need for text analytics in industries these days as it helps quantify the positive or negative perceptions of a company, brand, or product. Text analytics has a number of subdivisions like Information Extraction, Named Entity Recognition, and more broadly Natural Language Processing and Artificial Intelligence. The paper serves as a primer for text analytics.

[5] NLTK is a suite of program modules, data sets, tutorials and exercises, covering symbolic and statistical natural language processing. NLTK is written in Python and distributed under the GPL open source license. Over the past three years, NLTK has become popular in teaching and research.

[4] The authors Poon, Wu, Zhang decided to use a linear combination of all the successful method they had tried and they assigned weights to each model. After several iterations, their model resulted in a set of weights that gave them a low RMSE.

[6] Andrei Terentiev and Alanna Tempest found out that there is a correlation between the initial ten comments of a post and its score. Since they had a very small dataset, their results are not convincing enough and there may be other correlation factors.

IV. DESIGN

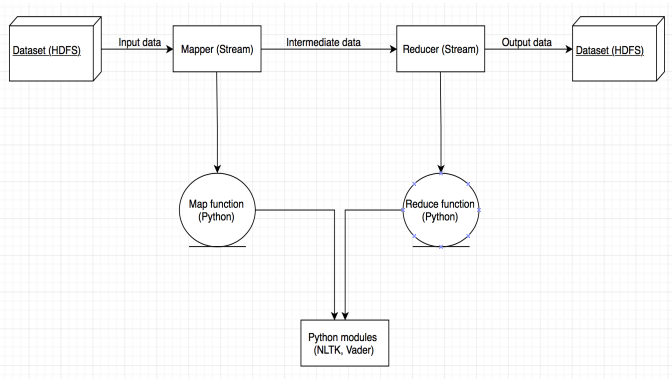

Fig 1

A. Dataset

Last year, Reddit released an enormous dataset containing all 1.7 billion of their publicly available comments. Kaggle hosts a portion of the comments (from May 2015). It is a csv file that has fields like subreddit_id, link_id, author, score, retrieved_on, controversiality, parent_id etc. We stored the dataset in HDFS and Hive.

B. Mapper/Reducer

We used both python's MapReduce framework and Hive to process the dataset. We used MapReduce for cleaning and profiling the data. We used Hive, which in turn generates a map-reduce job, to query the dataset interactively and derive some information about it.

C. NLTK

The Natural Language Toolkit is a machine learning based toolkit for the processing of human language data. We use NLTK library for python for the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging.

D. VADER

VADER (Valence Aware Dictionary and sentiment Reasoner) is a rule based sentiment analysis tool that we use to classify sentiment of the user comments into five classes – highly negative, moderately negative, neutral, moderately positive, highly positive.

V. RESULTS

We found some interesting trends while doing our analytic on the datasets released by Reddit(posts and comments) and Hackernews. Every post made in Reddit should be under a category, or, in Reddit's terms, Subreddit. When doing our analytic, we found that the posts made around 9 am EST usually achieves a high score. Participation begins to spike around that time for most of the subreddits except "nba" for which it spikes around 7 pm EST which coincides with the fact that nba games are scheduled at 7, 8 and 9 pm EST. We used engagement of a subreddit, sentiment of a post, and time of the day in which the post was made to make sense to our analytic.

Engagement:

In order to quantify the activities under each subreddit, we derived a new feature called engagement. It considers the number of links under each subreddit and the number of comments under each link, along with the number of authors participating in each subreddit. It indicates how engaging each subreddit is. By using this feature on the Reddit post dataset, we found the top 10 engaging subreddits and proceeded the analytic on the Reddit comment dataset.

Sentiment:

We then identified the sentiment of each comment and assigned them into 5 categories. Fig 2 shows the engagement trend for each subreddit across all the 5 sentiment categories. In subreddits like *news, todayIlearnt* the engagement in negative sentiment is high compared to other sentiments.

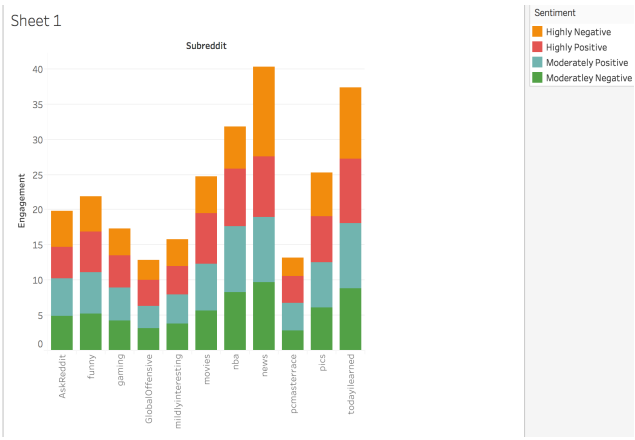

Fig 2

Fig 3 shows a similar distribution of sentiment engagement across each subreddit in a parallel bar graph

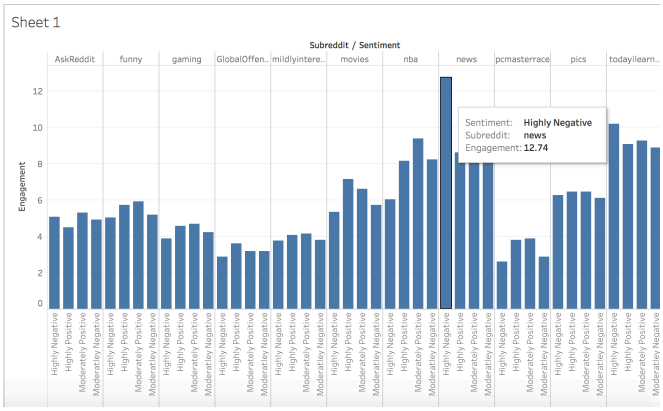

Fig 3

Timeline:
We classified the comments in each subreddit and sentiment class into 24 hour buckets and 31 day buckets. Fig 4 shows the timeline of comments made in each subreddit on any day.

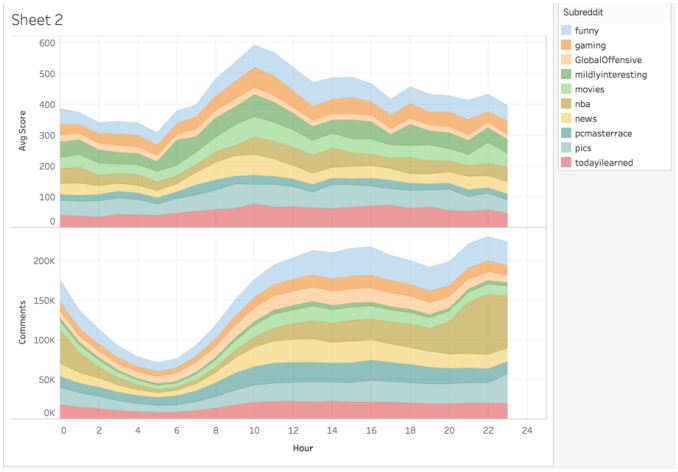

Fig 4

Ngrams:
We used the ngram library in python to find the number of occurences of every bigram and trigram in our dataset after removing the stopwords from the comments. We also found the average score that was achieved by the bigram and trigram in our dataset.

Prediction:
We had used 80% of our dataset to develop an analytic and tried our hands on the remaining 20% of the dataset to predict the influencing factors of a post. According to our analytic, three factors influence the popularity of a post – author, timeline-sentiment and ngrams. We had normalized the total scores achieved by every author in every subreddit for a sentiment class on a scale of -1 to 1 and the total scores in every subreddit for every sentiment in every hour on a scale of 0 to 1. The purpose of this normalization is to find the influence of an author and the influence of posting a comment with a particular sentiment on a given hour of the day. Say, an author with a normalized score of close to 1, his post has a higher chance of getting a good score. A post made on a given time with a given sentiment having a normalized score of close to 1 has a greater influence toward the post score. Fig 5 shows the influence of authors posting in the subreddits – 'funny' and 'news' with a negative sentiment class(4 and 5) toward the score of a post.

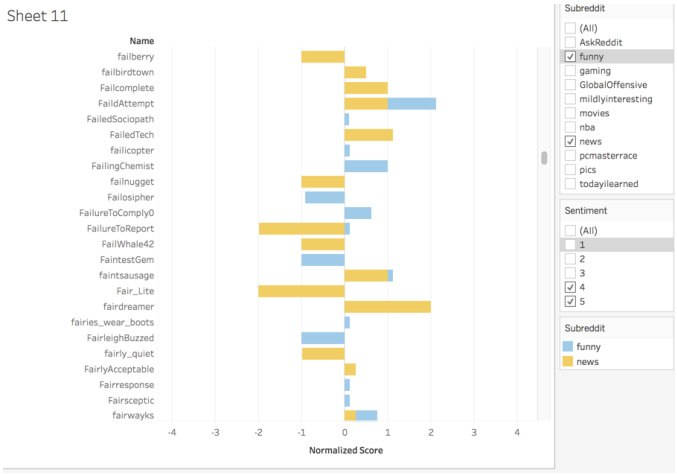

Fig 5

VI. FUTURE WORK

The title of the post appears in the homepage of reddit and hence plays a major role in getting the karma score. We were handicapped in that aspect as our dataset did not have the title of the post. One option was to use the reddit API to get the title of the posts for which our dataset had the comments, but it was not achievable in a limited period due to the constraints [2] set in the reddit API.

VII. CONCLUSION

We've identified 3 different features that conform to consistent patterns and influence the popularity of the posts and comments. Although this analytic gives predicitions of the popularity of a post to a desirable extent, some of the posts do not become popular unless they are posted multiple times. There has been widespread underprovision on social media wherein popular posts have been overlooked the first time they were posted. Factors as simple as the time of day in which the post was made could attribute to underprovisioning. [10]

VIII. ACKNOWLEDGMENT

We would like to thank Prof. Suzanne McIntosh for her patient advice, guidance and encouragment throughout the project. We were involved in scrum meetings on a weekly basis to discuss our progress and were given valuable feedback on how to proceed further.

IX. REFERENCES

[1] PREDICTING REDDIT POST POPULARITY by Jordan Segall and Alex Zamoshchin A. Gates. Programming Pig. O'Reilly Media Inc.,Sebastopol, CA, October 2011.

[2] PREDICTING REDDIT POST POPULARITY by Alex Bragdon, Ben McMorran, Himanshu Sahay, Lambert Wang.

[3] Text Analytics: the convergence of Big Data and Artificial Intelligence by Antonio Moreno1 , Teófilo Redondo2 https://dialnet.unirioja.es/descarga/articulo/5573981.pdf

[4] Reddit Recommendation System by Daniel Poon, Yu Wu, David (Qifan) Zhang CS229, Stanford University December 11th, 2011

[5] Multidisciplinary Instruction with the Natural Language Toolkit by Steven Bird, Ewan Klein, Edward Loper, Jason Baldridge

[6] PREDICTING REDDIT POST POPULARITY VIA INITIAL COMMENTARY by Andrei Terentiev and Alanna Tempest

[7] Telegraph (2013). ., [http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html]

[8] Facebook statistics (2013) [https://newsroom.fb.com/News]

[9] YouTube Statistics (2013). [http://www.youtube.com/yt/press/statistics.html

[10] Widespread Underprovision on Reddit by Eric Gilbert