Reddit comments analysis

Yash Choksi, Sweta Tripathi, Karolina Stachurska, Sowmya Mareedu, Sindhura Vemuri

Department of Information Systems, California State University

Los Angeles

e-mail : [ychoksi@caltstatela.edu](mailto:ychoksi@caltstatela.edu),

[stripat@calstatela.edu](mailto:stripat@calstatela.edu),

[kstachu@calstatela.edu](mailto:kstachu@calstatela.edu),

[smareed@calstatela.edu](mailto:smareed@calstatela.edu),

[svemuri@calstatela.edu](mailto:svemuri@calstatela.edu)

**1. Abstract**

Reddit a social networking website has so many comments on daily basics and we are analyzing those comments using hive query language. One of the important thing in today’s world is people are posting and saying so many things online so to make this internet safe for everybody we are analyzing who is posting which type of materials online. Reddit is vrery popular platform and in this platfomr there is so many stuff which is posted online like every second and it becomes very hard for any company to do analysis of this huge enormous content. So, we are analyzing whole content which is posted online and making visuliazations so that we can proceess them in accordingly.

**2. Introduction**

Reddit is very popular social networking website and content generated is enormous. Everyday content is really huge we can not process all of that on traditional systems because as we are analyzing this dataset it’s size is nearly 16GB with more than 15 columns and 2119905 rows. So we need specialized systems which can able to process this big data. We are using Hadoop distributed; mapreduce systems and specialized hive language to query and process this data.

As far as data is concerned our main goal is we can able to generate some good charts and visuals so stakeholders can able to know what type of comments and how platform can be improved. As more and more people spending time on social media it’s really task of social media provider that they can able to make it more safer. Dataset details is as follows:

URL:https://www.kaggle.com/reddit/reddit-comments-may-2015

Dataset size: 16GB

Cluster version: Hadoop 2.8.5-amzn-4

No of nodes: 3

HDFS Capacity: 147GB

CPU Speed: 2.20GHz

Storage: 678GB

Hive Version: Hive 2.3.5-amzn-1

3. Background

Reddit released it’s comments dataset on Kaggle and they want us to make visualization and charts about it. Theey are extremely concerned about all the hate crimes and other aspect of speech issues everyone is facing everyday. And due to such issues the reputation of social media is always a big issue. This dataset is having more than 1.7 Billion records but because of processing limitation they made it public only part of that dataset. Dataset is composed of May2015 dataset. And this dataset is composed of comments only. There are more than 20 columns and they all are a part of them and due to that there is huge concern around there. They wants us to create great visualization for stake holders and decision makers to make it more understandable.

4. Related Work

In hive data analysis there are so many phases to learn it out. One of the papers which I referred[2] was that they tried to do analysis of diabeetis data which was in Hadoop data warehousing scheme and they did analysis of those data. Because they did in that way they ran some queries on those dataset and they tried to build some predictive models using those analysis. Those predictive models were not came to the point because data was pretty old and it was not properly feed in.

As I always believe to take this data analysis project to forward we can do is make predictive models as we do not have any labels onn our data. But as of now why we used hive would be greaet question to answer. So as we refer to this paper[3] they also referred that hive is much better for data analysis and querying data. Because it is simple and more efficient at least in the manner of hadoop ecosystem. Because it’s simple in case of SQL like language and also it’s using Map-Reduce technology so that we can able to prove it easily. I really liked the conclusion which was driven by Laxmi Lydia.

I know as I am claiming that my dataset is really huge but in case of data warehousing and it’s related technologies we can easily say that this much datasize is nothing. As I read my third paper[4] in this they mentioned technologies that how they process data in warehousing because it’s always the case that data grows extremely faster and there are so many ways we have to process it. Now traditiona systems has no clue how to process it.

So, by looking at all the above mentioned approaches I would like to conclude that there is so much more things can be acgived with this approach and data size and complexity is increasing day-by-day.

5. Data Manipulation

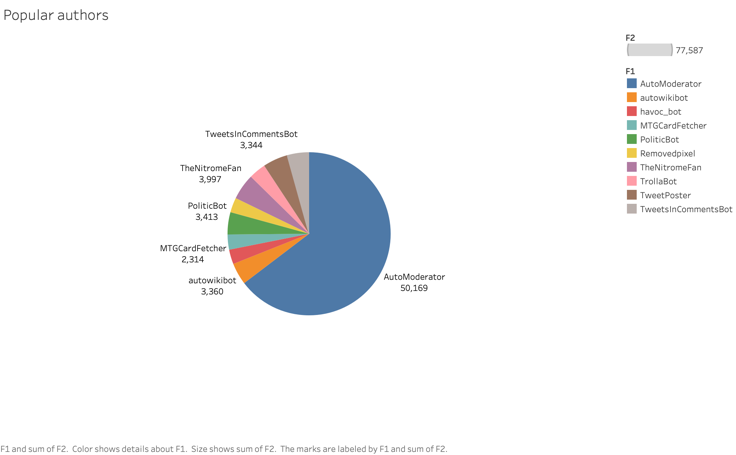
As data manipulation is main part of any project for us the process is somewhat same. As we get our data from Kaggele we gathered our data in the form of MySQL lite table. And this table size was 16 GB. Now to process this data in hive query language it’s really difficult to process this data in such enviourmennt. So we converted our dataset in CSV file format using SQL workbench. Using that software we gathered our data in csv format once we get that data in CSV format we uploded that data in HDFS file system which is an Oracle cluster.

That Oracle big data cluster is multi core computing enginee and using that engine we can able to compute using Map-Redcue on it. Because to process this hugee data on simple computer would be really hard and sometimes not even possible.

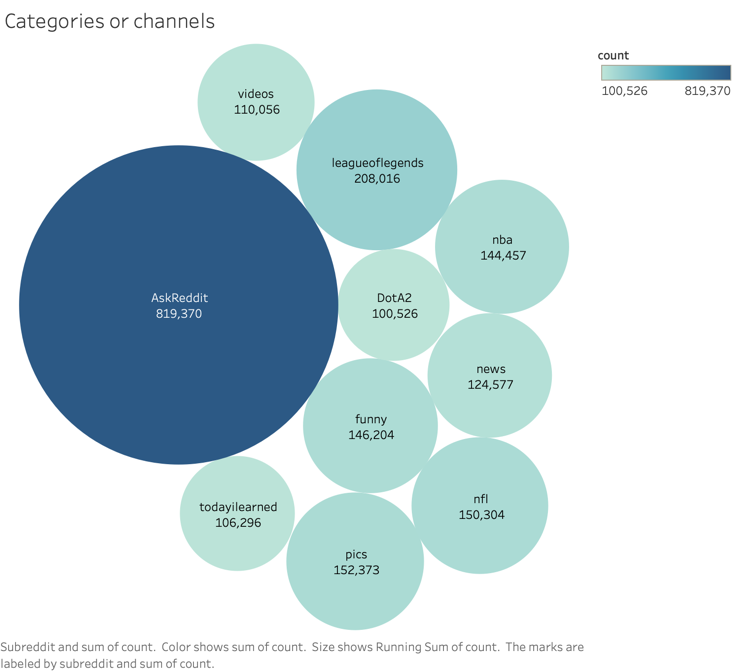
After uploading data over there we able to do is we created external table and fetch whole data in the table using hive queries and we created one hive table. Now, part comes is data cleaning data cleaning we removed all thee rows inn which author names are not present. The rest of dataset is still dirty but we try to remove all of them then there is nothing much left there to explore.

After doing all those process we queried data in different combinations and those dataset we found some good insights which we tried to show you here and those data we took it in different small csv tables.

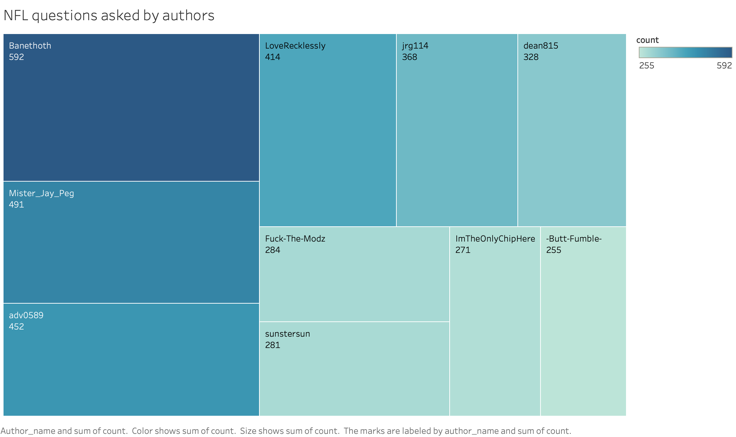
Those CSV tables we downloaded from cluster and those files we used to make different visualizations and we created all of those visualizations on Tableau. As we created visualizations every visualization contain it’s own meaning and some of the visualization are so vitall and meaningful that they represent user behavior and how they act on website such as on reddit or any other social media websites. We also worked on sentiment annalytis and we found that some users are very damaging and their languages are also very much offensive and damaging that it would be very difficult to even guide them to put in such direction. Then reddit made a rule of controversiality like if any user post something damaging then they should have to edit their post and make it less controversial. And as more edits required then those posts are more and more controversials.



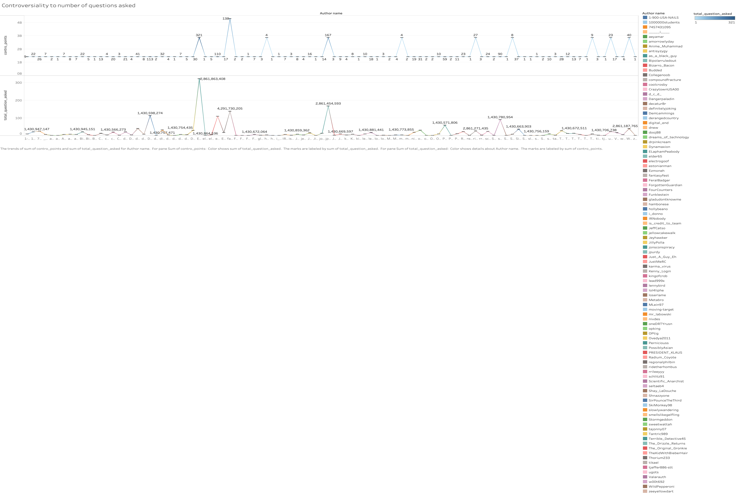
These are most popular authors and they represent by the number of questions they have posted on reddit.



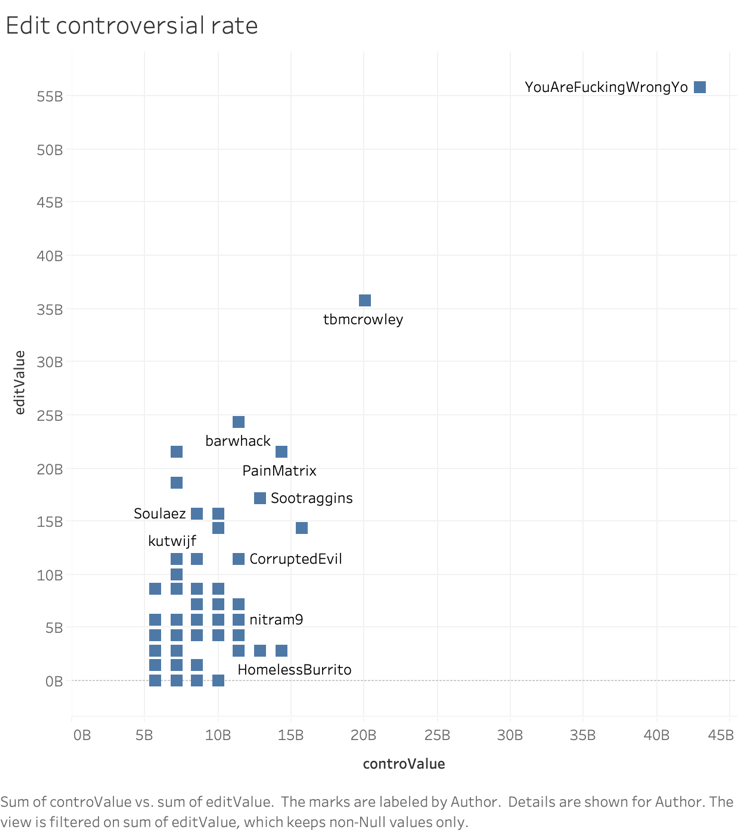
Reddit has created a good way to represent the different categories like here funny, nba and news and many more. So each category represent something and they are considereed as channels.



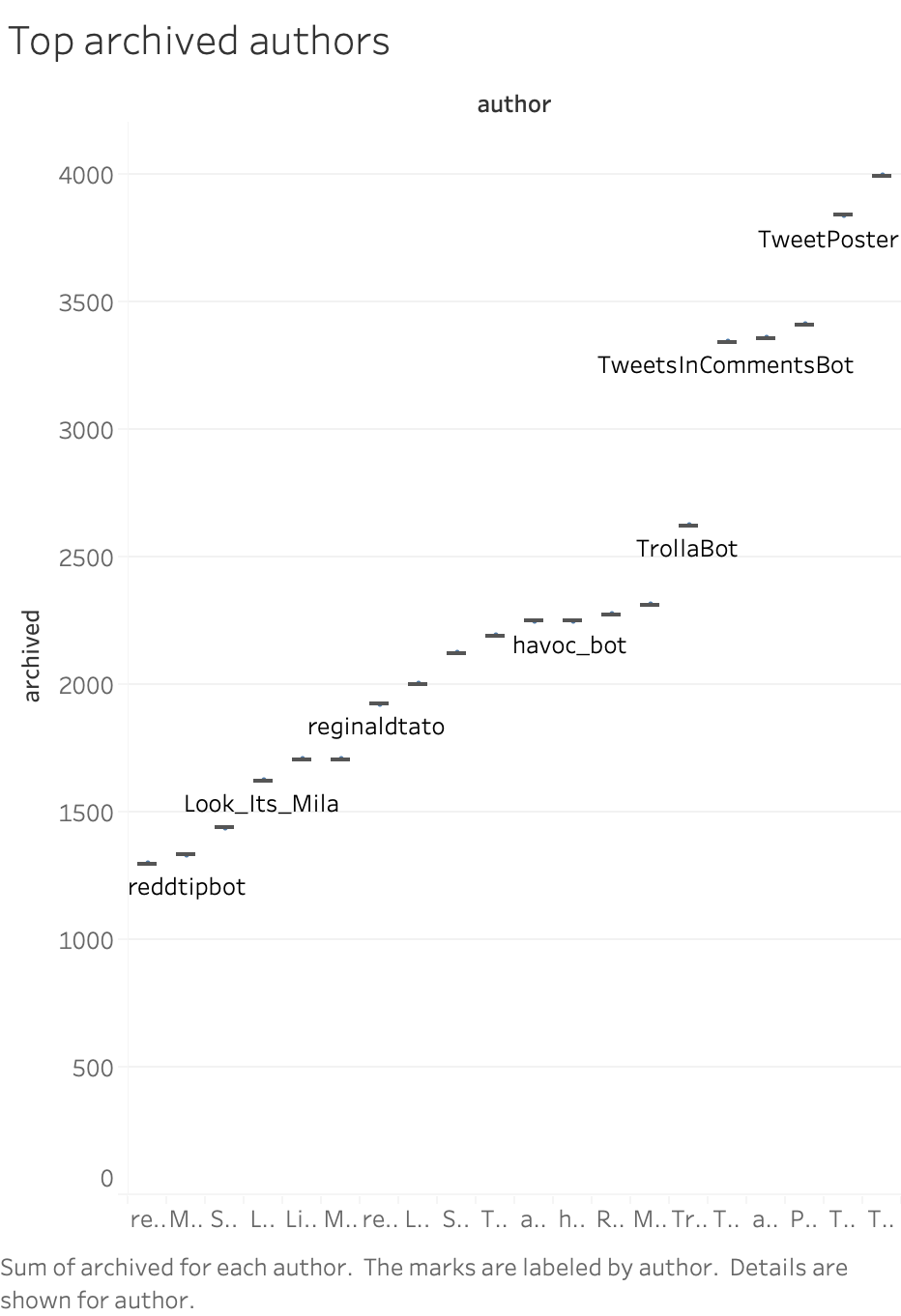
This visualization regers us to the idea of popuar authors who posted something on NFL only. Other categories does not taken in consideration. And by keeping that inn reference we can say that thesea are top 10 popular authors who posted something on NFL.



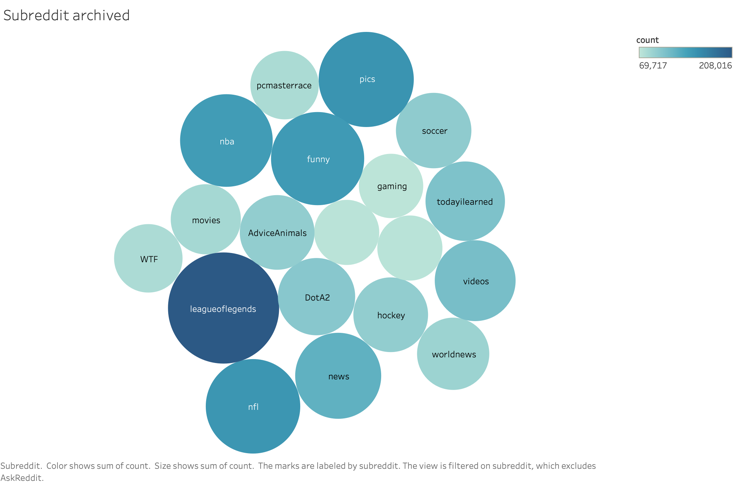
This is classical example of sentiment analysis here we tried to show that number of questions how much controversial questions is being asked by each authors and some asked only some questions but still their controversiality is so high.



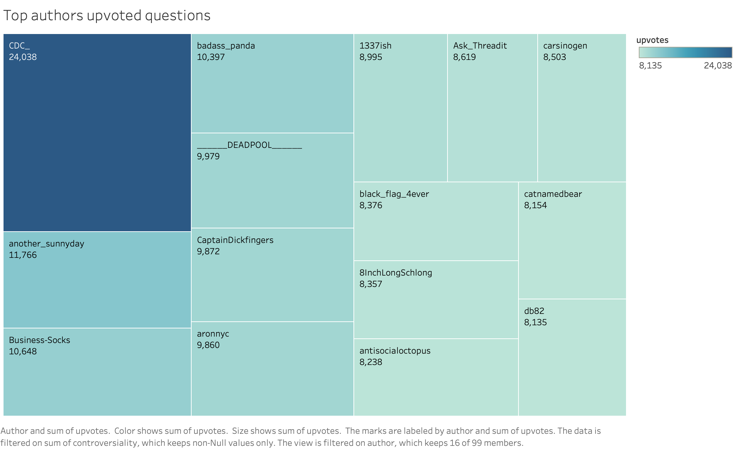
So reddit made a policy in which if your question is controversial then you have to edit it or make it non-controversial. So here we can see that as eeffect like some authors trieed to edit it but still their controversial rate is pretty high. And you can also see that as the controversiality is higher eedit rate is also higher.



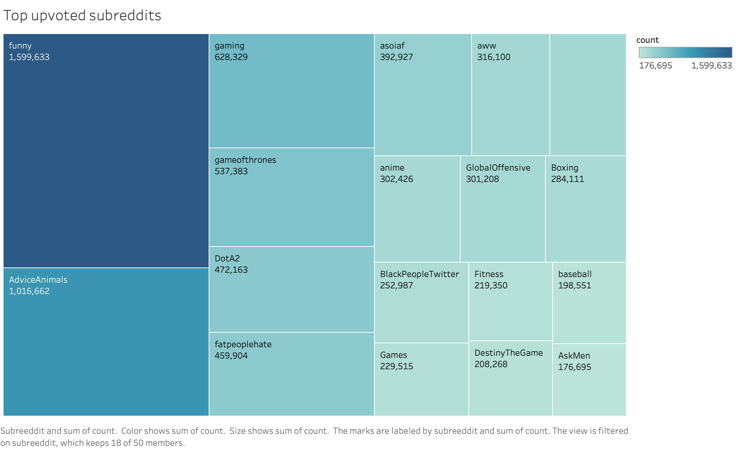
Reddit archives something only when it found that question has been answered and there is no more improvement is possible with the answer. Typically we say it as fact based answer. Also we need to understand that there are very few authors who ask on fact based questions. Only few authors aree there who ask fact based questions. So reddit should have to make strict rules in this scenario like stack oveerflow.



These bubble chart is for top channels which has maximum archived posts on the account. There are some of the channels which is archived most and some are not. Again rules are samee like channels which are archived most means the answers are preserved by reddit and there is no more scope of improvement in the answer; fact based answers.



Now always good questions are upvoted and there are some authors whose questions are upvoted and here are top authors whose questions are most upvoted. Like CDC\_ named author has most upvoted questions.



These are top upvoted channels as we have top upvoted subreddits. They are ran on hive quesry and we filtered most upvoted subreddits and these are most popular subjects on reddit.

6. Conclusion

In this project we analyzed sentiment analysis of different authors with different levels of controversial comments. One of th important findings we could figure out that some of the authors are posting so many comments but they are not that harmful for whole community, while some others just put some bunch of them and it’s very much damaging. We also try to analyze that some of the comments even are creating so much issues in day to day life and they are posted by such authors who should uinderstand such consequences.

**7. GitHub Link**

1. https://github.com/yashchks87/Reddit\_comments

**8. Refrences**

1.https://www.kaggle.com/reddit/reddit-comments-may-2015

2.http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.637.9347&rep=rep1&type=pdf

3.https://www.researchgate.net/publication/336148375\_A\_Hive\_and\_SQL\_Case\_Study\_in\_Cloud\_Data\_Analytics

4.http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.637.9347&rep=rep1&type=pdf

5.https://www.sciencedirect.com/science/article/pii/S2352914816000046