Uppsala University

Data Engineering Group 13 Change in usage of profanity in Reddit comments between 2006-2011

Author:
Joakim Michalak, Mandus Hjelm, Salman Khan, Şihab Sakıb Bayraktar, Eric Jonsson

March 24, 2021

1 Background

Usually in psychology physical tests are run to test hypothesis and theories regarding the human behaviour. But certain aspects of human behaviour can never be truly replicated in study groups and one such aspect is racism. In a group of people using any type of study may it be single blind or double blind, there is always an aspect of individuality when studying taboo topics like racism causing the final theory to be skewed in many certain ways due to these biases. Furthermore, if one wants to see a trend of every changing human behaviour, the tests need to be done in high frequency and with higher number of test subjects. However, seeing this is not truly possible, one can easily venture to the online community where anonymity and secrecy leads to people showing their true colors under pretext of veiled usernames. This is where human behaviour can be noted down and a trend can be captured against the same usernames under the popular and very candid platform called as Reddit. The same usernames' comments can be tracked down under certain subreddits to check how they changed over the years and how herd mentality affected the vicious nature of their comments and viewpoints. Certain studies have been done in the past as well such as "She's Reddit: A source of statistically significant gendered interest information?" [4] where authors Mike Thelwall and Emma Stuart show gender differences and discrimination by citing variations in interests and comments on Reddit. Furthermore, in "Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media" the authors Shan Jiang and Christo Wilson investigate linguistic signals[3] expressed in user comments in the presence of misinformation and fact-checking. They conclude how the language and demeanour of people changes when the topic being discussed has misinformation compared to topics that are presented on facts. Hence, it can be seen that even in virtual presence, our human psyche cannot be hidden from the real world and comments on platforms such as Reddit do advocate for us and our personalities.

With this background, our goal is to investigate if swearing is increasing in the comments from Reddit between the years 2006 and 2011.

2 Data format

The dataset is in JSON format that provides the keys: text comment, date, which sub-forum, etcetera, in ranges from 2006-2019. Since the dataset is JSON formated filtering to only leave the text string with the comment can easily be done to shrink the dataset in memory. JSON is a very suitable data format for this type of information since it is not static like a spreadsheet or SQL but not unordered like plain text.

3 Computational experiments

3.a Method and choice of framework

The choice of data storage using HDFS was because we wanted to be able to easily scale into all the available comments without using a VM with unnecessary size ref(Immutability Changes Everything). Since HDFS spread out the data in a cluster, more nodes/VMs can be added later if needed. HDFS works well with scalable frameworks MapReduce and Spark. The choice fell on Spark because of the possibility to read the information in to memory which will make analysis much faster[6].

In addition to faster analysis, we wanted flexibility and a framework that would support an interactive workflow. Setting up a spark cluster allowed us to focus on the infrastructure first and simplified the analysis as we would be able to test and try different implementations quickly by using a Jupyter Notebook. In contrast, had we used Hadoop MapReduce, there would be less configuration of the infrastructure but would result in a more rigid system when implementing the applications. Our main goal was to build a robust system for analysis, and not the analysis itself which further cemented our choice to use Spark over Hadoop MapReduce.

Our cluster consisted of three 4 core, 2gb RAM instances. On the SNIC cloud referred to as *medium* flavor. For our HDFS, we designated one node the name node and decided that this node would not also act as a datanode. This choice was made to keep the system divided into separate parts primarily for debugging purposes and avoiding adding complexity to the system too early. For our Spark cluster, we decided to use the master node as a worker as well. The driver was run on the master node, with forwarding set up to connect to the notebook through a networked computer.

3.b Scalability experiment

Weak and strong scaling are used as experiments in order to measure the scalability of the cluster. Weak scaling concerns a experiment in which the execution time is measured based on increasing problem size and number of cores simultaneously. This allows us to measure whether the execution time stays constant, or any overhead that the infrastructure may have, while also give a rough estimate of how much computing time would be needed for a certain application. Strong scaling only concerns execution time based on increasing the number of cores, but for a fixed problem size. Ideally, we want the execution time to decrease proportionally to how many cores are added. For example, doubling the core should ideally halve the execution time. 100% efficieny is not expected, however there should be a clear trend of decreasing execution time.

In figure 10 we can see plots of both strong and weak scaling for an analysis job of a 1gb file. For weak scaling, the 1gb file was analysed once per available core, meaning for 2 cores the 1gb was analysed twice, with 6 cores available the data was analyzed 6 times, and so on. This is not a fair comparison, as data locality and transfer speed can be optimized greatly. It also affects how much overhead affects performance.

3.b.1 Yarn

The resource manager (RM) and node manager work together to monitor resources and assist the scheduler in allocating resources [2]. There are many resource managers and selecting the right resource manager depends on many factors. Without a deep understanding of the framework, the cluster and the type of analysis that will be performed, selecting the right resource manager for the job is not a trivial task.

A few examples of resource managers are *Standalone*,*Hadoop Yarn* and *Apache Mesos*. Since both standalone and Yarn are available with little added configuration, we tested the scalability using both the standalone and Yarn resource manager. It is important to note that the location of data can affect performance as well, since the size of the data being analyzed here is relatively small, the real benefits of one resource manager over another may be hard to

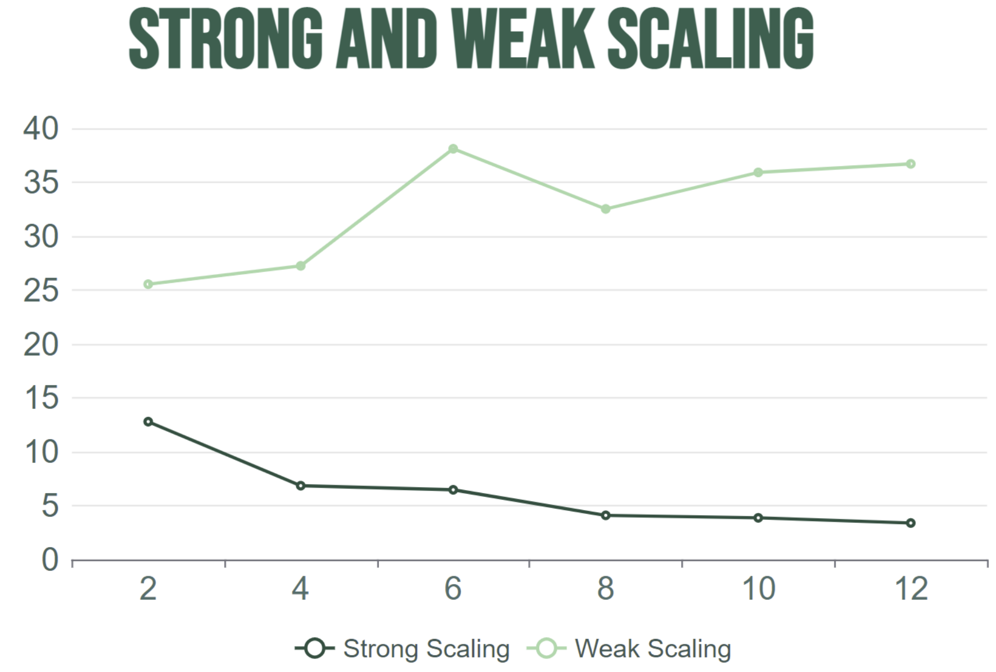

Figure 1: Scaling number of CPU cores with increasing data to process, and scaling with constant data in minutes.

identify. However that is also a reason to always experiment with different resource managers for different cluster setups and tasks. In figure 2 we plot weak scalability performance using both Yarn and Standalone, however, the results using Yarn suggest that not all cores were properly utilized as the time increases linearly with growing data.

3.b.2 Scalability results

While our experiments with Yarn proved unsuccessful and our scalability tests are simplified, figure **??** clearly show that our solution is horizontally scalable. As we increase the number of cores, we see a clear reduction in computation time. When the size of the dataset grows with the number of cores, I.E weak scalability, the computation time grows almost linearly. Again, the test is not accurate for a real world example, however we can be confident that the solution scales well given our tests.

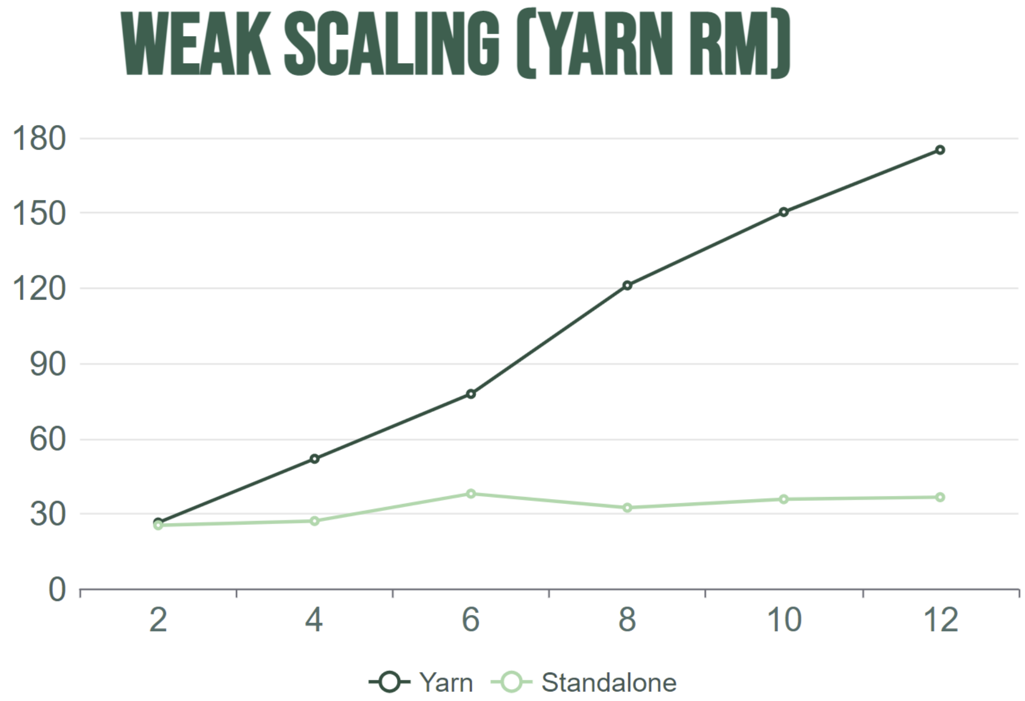

Figure 2: Comparing weak scalability using Yarn as resource manager vs Standalone resource manager.

4 Discussion and Conclusion

At the start, we raised the concern that human behaviour can be seen by online presence on Reddit. After going to the dataset and transforming them into various bar charts, it can be seen that a trend appears throughout the years. One can see how frequency of certain words rose throughout the years from 2006-2011 and one particular word decreased in frequency as well throughout these years.

From Figure 3 we can easily see how the usage of the word “fuck” has increased from 0.002 Normalized Count in 2006 to a hefty 0.01 Normalized Count in both 2010 and 2011. The percentage change from its lowest value and highest value is 400% which is the highest increase we have seen recorded in our dataset. This is a cause of concern showing how rise in ease of internet access has led to increase in people connected to it. As the number of people grew exponentially throughout the years, so did the usage of one of the most common negative word in the online internet community.

This can also be seen reflected on Figure 4 as well where the usage of word “motherfucker” increased from 0.08 in 2006 to 0.16 in 2010 and 0.14 in 2011 per 1000 comments. The percentage change between 2006 and 2011 is 75%. The trend for this particular word is quite different from the one in Figure 3 detailed above. One could speculate reason for such a varying trend might be due to the explicit obscenity associated with the word as well as various rules that would have been implemented to stop usage of this word. This is why even though there has been increase throughout the dataset, the pattern is not as quite obvious as other terms.

Figure 5 gives us one of the biggest changes when we draw conclusions for the considered period, against usage of the word “cunt”. The value in 2006 was 0 Normalized count whereas it rose to its highest value ever of 0.3 per 1000 comments in 2010 and then tapered off to 0.25 per 1000 comments in 2011. The dataset shows us how cultural linguistics can affect the online community. The etymology of the word “cunt” dates back to British English rather than American English. As internet brought people closer, it brought cultures closer as well and slang words intertwined. One of the main reasons why this word started from being non-existent online as per the dataset is due to two main reasons: the rise of British films winning awards and recognition in mainstream media in 2006 and 2007 and the fact that “cunt” was being used as a substitute of the word “fuck”. As per the documentary “The ’C’ Word” [1], it had become ”the most offensive insult one man could throw at another”.

To further show how insult words are substituted for other words, we will shine light on Figure 6 and Figure 7 together. The figures show how the usage of the word “darn” decreased for the years under question and how the usage of “damn” increased. This is due to how the word “damn” is known as the more intense and more explicit version of the word “darn”. As people online got to understand the anonymity behind online presence, they started to replace words that were “safe” to more explicit ones. These two figures show us exactly how the usage of word “darn” decreased from 0.3 to 0.1 per 1000 comments from 2006 to 2011 whereas the usage of word “damn” went from 2.8 in 2006 to 4.5 in 2011 per 1000 comments. The percentage change for the former is 67% decrease and the latter is 61% increase which is quite close showing correctness of theory stated above.

For Figure 8 we see a drastic rise of usage of word “asshole” from 2006 to 2007 (0.0004 to 0.0012). The value though remained quite close from 2007 till 2011 showing a similar pattern to the usage of word “fuck” that we saw in Figure 1. The percentage rise between the min and max points is 200% which is second highest increase we see after Figure 1.

For Figure 9 we see a straight increasing trend every year from 2006 till 2011 concerning the word “bitch”. One of the main reasons for this could be the increase of female presence in the online community and the inclusion of using insult words that are towards a certain group. The value in 2006 was 0.5 per 1000 comments and value in 2011 was 1.6 per 1000 comments. Which shows overall increase of 200%. However the most special thing about this figure and dataset is that from 2007 to 2011, each year there has been a steady increase of average 15%.

For the last figure, Figure 10, we see how insults which are specific for a certain minority and marginalized group has increased in the years under question. The word “faggot” is used as a derogatory term to insult a person belonging to the LGBTQ+ especially gay people. As there was an increase in the presence of LGBTQ community on the internet and increasing concern for them [5], so was the increase of slurs that are directed at them online under the pretext of anonymity. The highest increase was in 2010 from the previous year where it jumped from 0.08 to 0.16 per 1000 comments, doubling in usage. This can be because in 2009 and 2010 many countries and states in US started legalizing same-sex marriages which would have cause the spike in between these two years as it usually takes time for real-life events to translate on online communities such as Reddit.

In conclusion, it can be seen that the human psyche is able to run free and candidly on the internet. Platforms such as Reddit that usually do not have policing outlook on words and carry the banner of free-speech usually end up with online personalities that insult and abuse without remorse. From words that are universal to words that are made specifically to target marginalised minorities, the dataset has proven that they have increased for the years in question.

5 Authors Contribution

- Joakim Michalak, Mandus Hjelm, Salman Khan, Şihab Sakıb Bayraktar and Eric Jonsson contributed to *Conceptualization* and *Project Administration.* As well as the initial setup of the HDFS / Spark cluster.
- Joakim Michalak and Mandus Hjelm implemented the Spark Application and maintained HDFS data.
- Şihab Sakıb Bayraktar, Salman Khan and Eric Jonsson setup additional nodes to the cluster.

6 Appendix

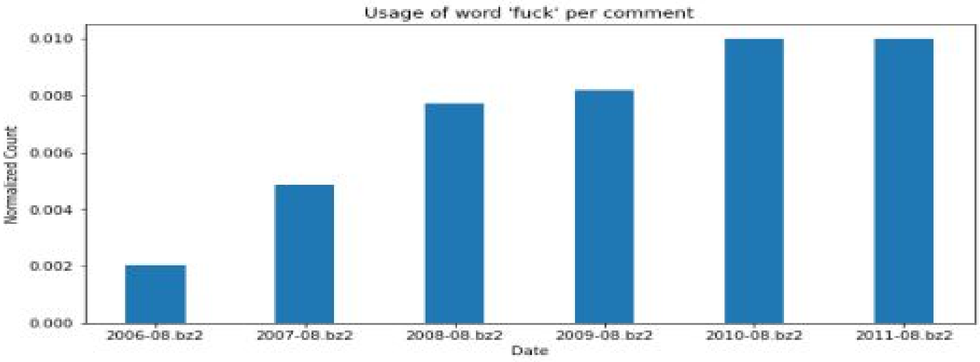

Figure 3: Showing the count of the word FUCK per 1000 comments used from 2006 to 2011

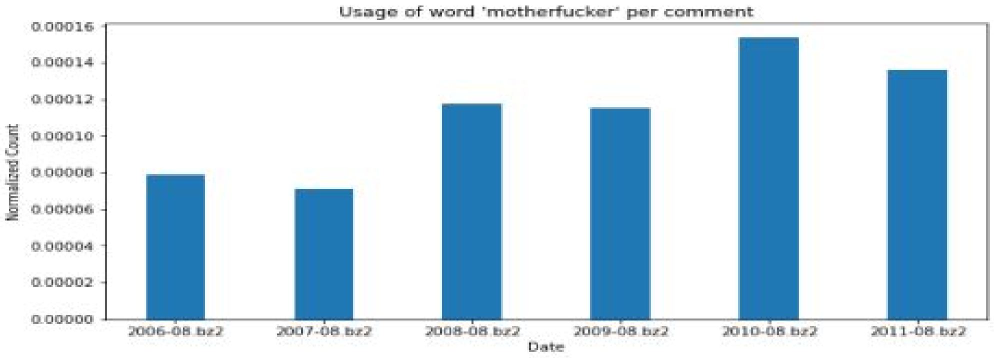

Figure 4: Showing the count of the word Mother Fucker per 1000 comments used from 2006 to 2011

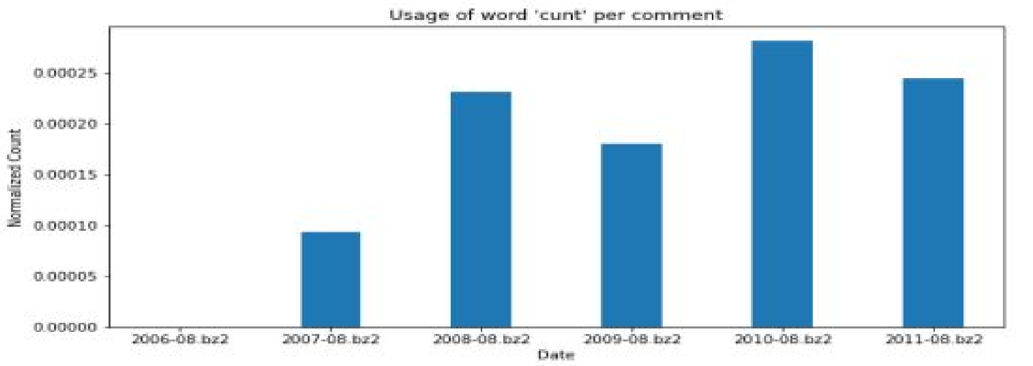

Figure 5: Showing the count of the word CUNT per 1000 comments used from 2006 to 2011

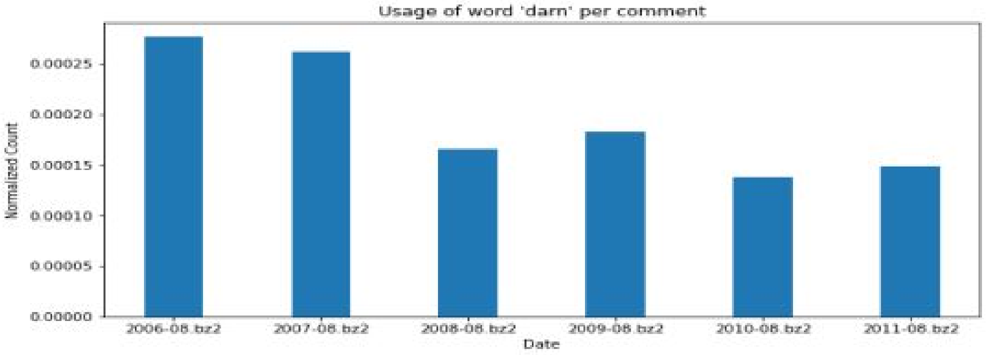

Figure 6: Showing the count of the word DARN per 1000 comments used from 2006 to 2011

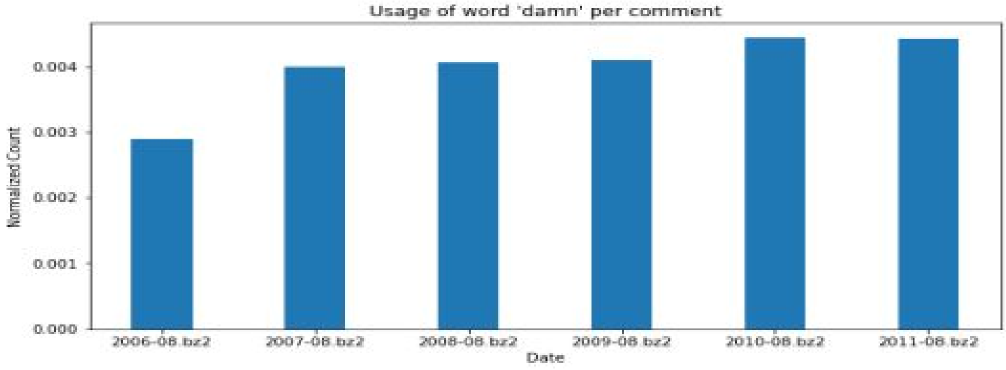

Figure 7: Showing the count of the word DAMN per 1000 comments used from 2006 to 2011

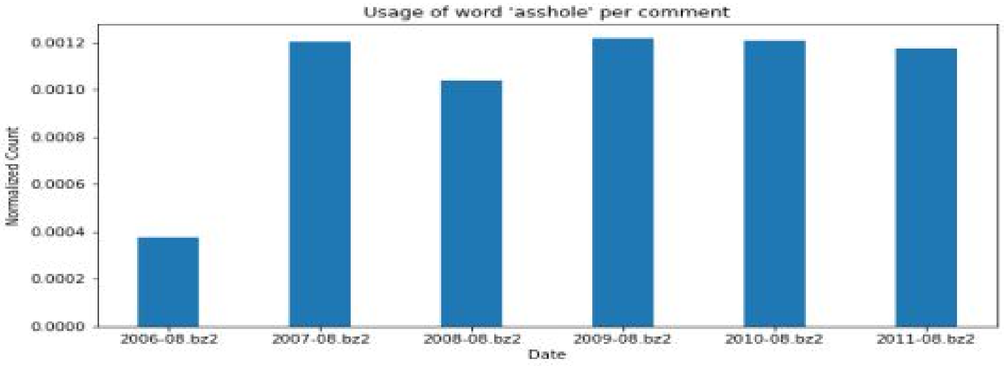

Figure 8: Showing the count of the word ASSHOLE per 1000 comments used from 2006 to 2011

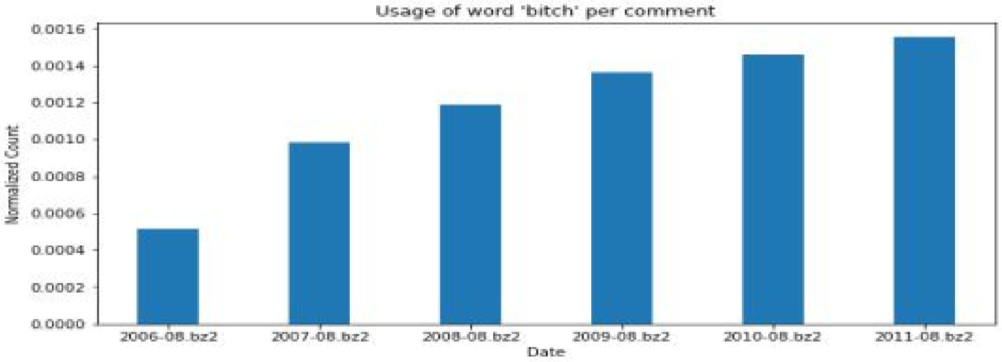

Figure 9: Showing the count of the word BITCH per 1000 comments used from 2006 to 2011

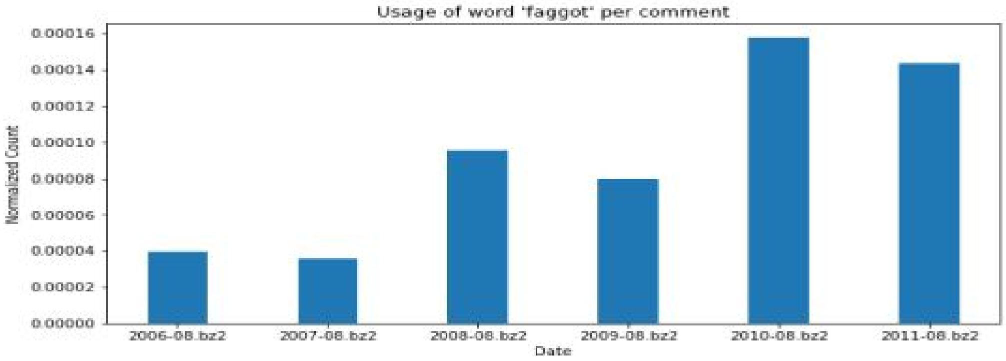

Figure 10: Showing the count of the word FAGGOT per 1000 comments used from 2006 to 2011

7 References

[1] M. David Z Hirsch. The c word. 2019.

[2] A. Hadoop. Apache hadoop yarn. *hadoop.apache.org*, 2020.

[3] S. Jiang and C. Wilson. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23, 2018.

[4] M. Thelwall and E. Stuart. She's reddit: A source of statistically significant gendered interest information? *Information processing & management*, 56(4):1543–1558, 2019.

[5] WikiPedia. Lgbt+ history. 2019.

[6] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. *University of California, Berkeley*, 2012.