Facilitating Analysis of Big Data on Reddit via an Easy to Use Visualisation Tool

Jorge Goncalves
University of Melbourne
j.goncalves@unimelb.edu.au

Simon Klakegg
University of Oulu
simon.klakegg@oulu.fi

Niels van Berkel
University of Melbourne
n.vanberkel@unimelb.edu.au

Simo Hosio
University of Oulu
simo.hosio@oulu.fi

With the rapid proliferation of social media sites, researchers have increasingly turned to data generated from these platforms to investigate human behaviour. In this paper we report the design and implementation of the RDV (Reddit Data Visualisation) platform, a visualisation tool aimed at facilitating the analysis of a publicly available Reddit dataset, which contains ~1.7 billion JSON objects collected from October 2007 to October 2015. RDV allows for researchers without advanced coding skills to easily analyse this dataset, while also providing a tailor-made platform to account for the intricacies of any dataset originating from Reddit. We showcase the features of the platform through an example of data analysis using the Reddit dataset: the 2015 United Kingdom general elections. Finally, we conclude by discussing the need for better and simpler visualisation tools for non-technical researchers to analyse Big Online Behavioural Datasets, and report our ongoing work in this area.

Social Media, Online Human Behaviour, Data Visualisation, Reddit, Big Data

1. INTRODUCTION

Millions of people login daily to several different internet platforms (*e.g.*, Facebook, Google, Twitter, Instagram and Reddit), which has fundamentally changed how humans communicate, seek information, discuss a plethora of topics, follow their interests, as well as a number of other activities. As a direct result of the widespread adoption of these platforms, the stream of user information being generated online has increased exponentially in the last few years. For this reason, researchers of different scientific domains have increasingly leveraged this data to investigate human behaviour. For instance, previous work has highlighted how the information contained within these social streams can provide rich insights on people's opinions and perceptions on a number of different topics [23,34], what motivates certain online behaviours [22,27], and effects of social networking sites on users [20]. However, an important challenge that scientists face with such large behavioural datasets is that of information overload due to the sheer quantity of available data [19].

In this paper we describe the design and implementation of a web-based platform, called RDV (Reddit Data Visualisation), meant to provide an easy and approachable way to analyse Reddit's entire publicly available comment dataset. Reddit is an entertainment, social networking, and news website where registered community members can submit content, such as text posts or direct links, making it essentially an online bulletin board system. The dataset consists of ~1.7 billion JSON objects complete with the comment body, score (including up and down votes), author, subreddit, position in comment tree, creation time and other fields that are available through Reddit's API. We created our platform by implementing a user-friendly web interface for making queries, and then presenting the information visually in the form of multiple charts such as bar charts, line charts, and word clouds. Through this platform we give interested Reddit users and researchers the possibility to easily analyse this rich dataset regarding different topics of their choosing.

We showcase changes in discussion points throughout the 2015 United Kingdom general election race. Finally, we discuss the need for better and simpler visualisation tools for non-technical researchers to analyse Big Online Behavioural Datasets, and report ongoing work aimed at improving the platform.

2. RELATED WORK

2.1 Analysing Online Human Behaviour

Emerging online communication technologies are fundamentally changing the way we behave, interact, and socialise [25]. It has been estimated that in 2016 over 2.3 billion people use at least one social media platform [5]. In recent years,

researchers have increasingly turned to this unprecedented source of data to investigate and better understand a plethora of different human behaviours. For instance, previous work has investigated if a user's social network structure can provide insights regarding his or her personality characteristics. As an example, previous work [30] found a relationship between social network structure on Facebook and social capital, and how this relationship is moderated by personality traits. In another study, researchers demonstrate through a network science approach that empathy is closely linked with social capital [31]. Social network analysis has also been used as proxy for studying empathy [33], which showed that empathy is mirrored in the structure of social ties among adolescents in German schools.

A large body of work has explored user behaviour towards privacy when using online social media. This is an important element of social media platforms, even though there is often a "privacy paradox", *i.e.* a discrepancy between people's privacy attitudes towards sharing information and their actual sharing patterns. For instance, previous work has revealed a high discrepancy between stated concerns and actual behaviour towards sharing static profile information [2,32]. Several techniques have been proposed to assist users in dealing with these privacy concerns, such as selective sharing [17] or narrowcasting [9,10]. Other work has investigated how social media data can be used to predict events, trends, and user reactions. For instance, researchers developed a framework that successfully predicted most of the new popular fashion models that appeared in 2015 based on images collected from Instagram [23]. Others have investigated how various socio-linguistic properties are responsible for hashtag compound formation on Twitter and propose a model to predict popular hashtag compounds [22]. There has also been work on modelling the competition dynamics that shape the fate of Facebook posts, and provided actionable insights to improve user engagement [11].

The data generated in these online platforms has also enabled large-scale studies linking lifestyle and health data at an individual and community level [8]. On the individual level, researchers developed a research framework that utilises the LIWC (Linguistic Inquiry and Word Count) to detect the real-time mood of Twitter users [24]. They found a correlation between the depressive state and the tweet sentiment of that user. Similarly, previous work has highlighted the potential to use social media to detect and diagnose major depressive disorder in individuals [7]. Specifically, they find that everyday social media use can leveraged to predict the onset of depression in individuals, as measured through decrease in social activity or raised negative affect. Other medical conditions have been studied by leveraging social media data, such as sleep problems, substance abuse and eating disorders. On the community level, researchers used smile recognition on 9 million geotagged Twitter images and developed a Smile Index as a formalised measure of societal happiness [1]. Further, Sadilek & Kautz show that the health of a population can be predicted based on their Twitter usage when coupled with other factors [26]. Next, we look at some of the visualisation tools for large datasets reported in literature.

2.2 Big Data Visualisation Tools

The appropriate visualisation of information contained within large datasets is an important challenge for analysts and researchers. Without tools to adequately explore the large quantities of information being collected, and despite its potential usefulness, the data becomes useless [18]. Tinati & Halford also consider the methodological challenges for those who engage with Big Data, and propose tools that address both the ephemeral, changing nature of platforms such as Twitter and the many temporary or permanent networks that form within one platform [29]. Further, inadequate visualisation of the information can lead to the researcher or data analyst to miss potential biases within the dataset. For example, demographic bias [13] and geographic bias [12] can have a significant impact in the produced outputs of the analysis meaning that decision-makers can ultimately be ill-informed.

There are several visualisation tools aimed at large datasets reported in literature. For instance, DEVise is a data exploration system that allows users to easily develop, browse, and share visual presentations of large tabular datasets from several sources [21]. Another example is ParaView, a tool that provides a graphical user interface for the creation and dynamic execution of visualisation tasks. It supports the visualisation and rendering of large datasets by executing these programs in parallel [3]. Dendroscope is another tool aimed visualising and navigating both small and large datasets, specifically phylogenetic trees [15]. More recently, an open source visualisation tool called VisIt was developed for visualising and analysing particularly large datasets. VisIt is aimed at enabling data understanding, scalable support for growing data, and providing a robust and usable product for end users [6].

While there are numerous other examples of such visualisation tools, many of them either require advanced coding skills or do not adapt well to certain datasets as they were developed with broad use in mind. The platform reported here is specifically tailored to the intricacies of the Reddit dataset. This enables interested researchers to easily pick up the platform as there is a reduced learning and configuration barrier. We describe the dataset and the platform in more detail in the following sections.

3. REDDIT DATASET

As of July 2016, Reddit is the 26th most visited website in the world with over 500 million monthly visitors [28]. The dataset consists of over 1.7 billion JSON objects starting from October 2007 up until October 2015. It contains several fields obtained through the Reddit's API including but not limited to: **id -** A unique identification number, **body -** Comment written by a Reddit user, **score -** Number of upvotes minus the number of downvotes, **upvotes:** Number of upvotes given to a particular comment, **downvotes:** Number of downvotes given to a particular comment, **author name:** Reddit user that wrote the comment, **subreddit:** Area of interest in which the comment was posted, **position in comment tree:** Parent comment or a response to another comment, **creation time:** Date and time the comment was posted, **gilded:** States if the comment was made by a user with Reddit gold.

4. THE RDV PLATFORM

RDV was implemented using the Shiny web framework with additional libraries for both the front end and the back end of the web application. Shiny Dashboard is a package for R that provides a set of functions designed to create analyses into interactive web applications. It is easy to use and configure making it ideal for analysing complex datasets, such as the one described here. Shiny Dashboard also allows dynamic and responsive design meaning a high-level of compatibility across multiple devices. The web-based platform is connected to a MySQL database where the Reddit dataset is stored. The database is stored on a large commercial server and the functional web page on a smaller local server. In its current iteration, the platform has three different analysis patterns, which will generate varying data visualisations based on selection parameters defined by the user:

- **amount of comments:** provides visualisation of the frequency of comments through bar or line charts.
- **subreddit relations:** provides visualisation of relationships between subreddits based on certain threshold of shared commenters through a network graph.
- **frequency of words:** provides visualisation of word frequency through word clouds. Word clouds give greater prominence to words that appear more frequently within the selected data and ignores stopwords (*e.g.*, the, a, be, as, for).

There are several features that are common across the three patterns, including:

1) selecting the time period for which to perform the analysis,
2) limiting the analysis to certain subreddits (*e.g.*, look at comments made only in the Politics subreddit),
3) defining the comments' score range (*i.e.*, selecting only comments with a certain score range) and
4) selecting comments' gilded status (*i.e.*, select all comments, only gilded comment or only non-gilded comments).

In addition, each pattern has specific features to allow for more control from the user. In the amount of comments pattern, the user can also define a list of keywords (*i.e.* only comments with certain words will be taken into account and a list of authors (*i.e.* only comments from certain authors will be taken into account). In the subreddit relations pattern the user can also define the percentage of shared commenters and the minimum size of the subreddits included in the analysis. Finally, the frequency of words pattern allows the user to set the minimum frequency for a word to be considered and the maximum number of words that appear in the cloud.

After the user presses the "Plot" button, the platform will then process the request and generate the plot. Depending on the scope of the analysis, this can take between a few seconds to several hours (*e.g.*, if the whole the dataset is selected). In Figure 1, we can see a plot generated from the amount of comments pattern that shows frequency of comments throughout 16 days.

Figure 1: *Plot overview.*

On the right, the user can change the plot type (in this case the options are bar or line) as well as an option to separate the visualisation between different selected subreddits. This is achieved through stacked bar plots or different coloured lines in a line plot to facilitate visualising the differences between the selected subreddits. The user can also save the generated plot in a format of their choosing (JPG, PNG, PDF). Finally, in the additional information section the user can check the query as well as the returned rows based on their selection

criteria by clicking the dropdown button. This allows the user to more closely inspect the returned output directly on the platform (includes a search feature) or extract it for further analysis (*e.g.* content analysis) if the user wishes to do so.

5. EXAMPLE OF ANALYSIS BASED ON RDV

5.1 The 2015 United Kingdom General Election

The United Kingdom general election of 2015 took place on the 7th of May 2015, which elected the 56th Parliament of the United Kingdom. The election aimed at appointing one Member of Parliament to the House of Commons from each of the 650 parliamentary constituencies. The Conservatives won a 12-seat majority in parliament with David Cameron being re-elected Prime Minister of the United Kingdom. We analysed the frequency of comments in the UKPolitics subreddit, which is the most subscribed subreddit regarding British Politics. Figure 2 shows the difference in number of comments posted on the day of the vote as well as the days of the preceding and subsequent week. There was higher activity within the subreddit on the day after the election, as at that point all the votes had been counted and users began to react to the outcome.

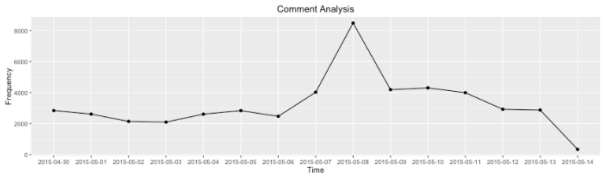

Figure 2: *Frequency of comments made in the UKPolitics subreddit.*

We then created two different word clouds using comments from the UKPolitics subreddit, one during a period leading up to the election (1st January 2015 - 7th May 2015, Figure 3 left) and the other during a period after the election (8th May 2015 - 30th September 2015, Figure 3 right). The word clouds were constructed using only words that appeared at least 50 times.

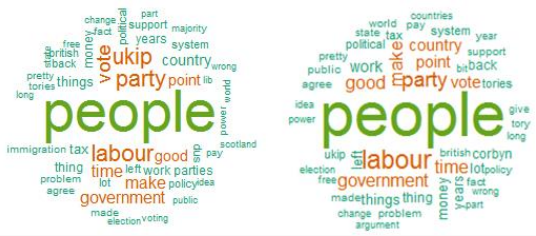

Figure 3: *Left: Word cloud between 1st of January 2015 and 7th of May 2015. Right: Word cloud between 8th of May 2015 and 30th of September 2015.*

Unsurprisingly, there are far more mentions of the Labour Party when compared to the Conservative Party (Tories), even though the latter won the election. The main Labour Party subreddit has over 4500 subscribers, while the main Conservative Party has around 600, which again highlights how the Reddit user base tends to lean liberal. During the period leading up to the election there were several issues being discussed by members of the community, namely immigration and taxes/money, which were the top two issues that decided the general election vote [16]. Post-election there was increase discussion regarding Jeremy Corbyn who took over the Labour party on the 12th September. Furthermore, content analysis of selected comments, which we intend to make possible directly on RDV in the future, can provide further insights on the community's opinion on different aspects of this event.

6. CONCLUSION AND ONGOING WORK

We present the design and implementation of the RDV platform, a visualisation tool aimed at facilitating the analysis of a ~1.7 million JSON object dataset from the Reddit social media platform. Previous work has emphasised the increased need of actionable insights when analysing social media for political and policy reasons [11]. By providing simpler visualisation tools that do not require advanced coding skills, a larger number of policy and politics researchers can analyse publicly available social media datasets to extract useful insights. Further, more specific and tailor-made tools to a certain social media generated dataset, such as the one presented here, enables a more in-depth analysis that in turn can lead to richer insights. Here, we showcase a couple of examples of analysis that can easily be achieved through the use of RDV, which can easily be extended to other scenarios.

We intend to continue developing our platform to provide additional features to enable a more in-depth and varied analysis of Reddit datasets. For instance, allowing the analyst or researcher to select only comments with a certain minimum number of votes. This would facilitate the identification of controversial comments made within a predefined subreddit. Another planned feature is to allow for content analysis of comments directly on the platform. An example of a potential use case of this feature would be conducting sentiment analysis of comments within certain subreddits or of certain members of the community. This would help identify and measure the opinions of users to events, policy changes, among others. Finally, we are interested in moving beyond Reddit and create tailor-made visualisation tools for other popular social media platform, such as Twitter.

8. REFERENCES

[1] Abdullah, S., Murnane, E.L., Costa, J.M., and Choudhury, T. 2015. Collective Smile: Measuring Societal Happiness from Geolocated Images. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, 361-374.

[2] Acquisti, A., and Gross, R. 2006. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. In Privacy Enhancing Technologies. Springer Berlin Heidelberg, 36-58.

[3] Ahrens, J., Geveci, B., Law, C., Hansen, C.D., and Johnson, C.R. 2005. ParaView: An End-User Tool for Large-Data Visualization. In Visualization Handbook, 717-731.

[4] Barthel, M. 2016. How the 2016 presidential campaign is being discussed on Reddit | Pew Research Center. Available at: http://www.pewresearch.org/fact-tank/2016/05/26/how-the-2016-presidential-campaign-is-being-discussed-on-reddit/ [Accessed August 22, 2017].

[5] Chaffey, D. 2016. Global social media research summary 2016. Smart Insights: Social Media Marketing.

[6] Childs, H. 2013. VisIt: An end-user tool for visualizing and analyzing very large data. University of California Technical Report.

[7] De Choudhury, M., Counts, S., Horvitz, E., and Gamon, M. 2013. Predicting Depression via Social Media. In AAAI Conference on Weblogs and Social Media, 2-11.

[8] Garimella, V.R.K., Alfayad, A., and Weber, I. 2016. Social Media Image Analysis for Public Health. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, 5543-5547.

[9] Goncalves, J. 2011. Groupster: Narrowcasting on Social Networking Sites. Master's Thesis, University of Madeira.

[10] Goncalves, J., Kostakos, V., and Venkatanathan, J. 2013. Narrowcasting in Social Media: Effects and Perceptions. In International Conference on Advances in Social Network Analysis and Mining, IEEE/ACM, 502-509.

[11] Goncalves, J., Liu, Y., Xiao, B., Chaundhry, S., Hosio, S. and Kostakos, V. 2015. Increasing the Reach of Government Social Media: A Case Study in Modeling Government-Citizen Interaction on Facebook. Policy & Internet, 7(1), 80-102.

[12] Grinberg, N., Naaman, M., Shaw, B., and Lotan, G., 2013. Extracting Diurnal Patterns of Real World Activity from Social Media. In ICWSM, AAAI.

[13] Harford, T. 2014. Big data: A big mistake? Significance, 11(5), p14-19.

[14] Hook, J. 2015. Support for Gay Marriage Hits All-Time High — WSJ/NBC News Poll. Available at: http://blogs.wsj.com/washwire/2015/03/09/support-for-gay-marriage-hits-all-time-high-wsjnbc-news-poll/ [Accessed August 17, 2017].

[15] Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M., and Rupp, R. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics, 8(1), 1.

[16] Ipsos MORI 2016. Economy, immigration and healthcare are Britons' top three issues deciding general election vote. Available at: https://www.ipsos-mori.com/researchpublications/researcharchive/3447/Economy-immigration-and-healthcare-are-Britons-top-three-issues-deciding-general-election-vote.aspx [Accessed August 22, 2017].

[17] Kairam, S., Brzozowski, M., Huffaker, D., and Chi, E. 2012. Talking in Circles: Selective Sharing in Google+. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 1065-1074.

[18] Keim, D.A. 2001. Visual Exploration of Large Data Sets. Commun. ACM, 44(8), 38-44.

[19] Kovach, B., and Rosenstiel, T. 2011. Blur: How to know what's true in the age of information overload, Bloomsbury Publishing USA.

[20] Liu, Y., Venkatanathan, J., Goncalves, J., Karapanos, E., and Kostakos, V., 2014. Modeling what friendship patterns on Facebook reveal about personality and social capital. ACM Transactions on Computer-Human Interaction, 21(3), 17:1-17:20.

[21] Livny, M., Ramakrishnan, R., Beyer, K., Chen, G., Donjerkovic, D., Lawande, S., Myllymaki, J., and Wenger, K. 1997. DEVise: Integrated Querying and Visual Exploration of Large Datasets. SIGMOD Rec, 26(2), 301-312.

[22] Maity, S.K., Saraf, R. and Mukherjee, A., 2016. #Bieber + #Blast = #BieberBlast: Early Prediction of Popular Hashtag Compounds. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, ACM, 50-63.

[23] Park, J., Ciampaglia, G.L., and Ferrara, E. 2016. Style in the Age of Instagram: Predicting Success Within the Fashion Industry Using Social Media. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, ACM, 64-73.

[24] Park, M., Cha, C., and Cha, M. 2012. Depressive Moods of Users Portrayed in Twitter. In KDD Workshop on Health Informatics, ACM, 1-8.

[25] Reynolds, B., Venkatanathan, J., Goncalves, J., and Kostakos, V. 2011. Sharing ephemeral information in online social networks: privacy perceptions and behaviours. In IFIP TC.13 International Conference on Human-Computer Interaction, Springer, 204-215.

[26] Sadilek, A., and Kautz, H. 2013. Modeling the Impact of Lifestyle on Health at Scale. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, ACM, 637-646.

[27] Schrammel, J., Köffel, C., and Tscheligi, M. 2009. How Much Do You Tell?: Information Disclosure Behaviour Indifferent Types of Online Communities. In Proceedings of the Fourth International Conference on Communities and Technologies, ACM, 275-284.

[28] SimilarWeb. 2017. Available at: https://www.similarweb.com/website/reddit.com [Accessed August 5, 2017].

[29] Tinati, R., and Halford, S., 2012. Interrogating Big Data for Social Scientific Research-An Analytical Platform for Visualising Twitter.

[30] Venkatanathan, J., Karapanos, E., Kostakos, V., and Goncalves, J. 2012. Network, personality and social capital. In ACM Web Science, ACM, 326-329.

[31] Venkatanathan, J., Karapanos, E., Kostakos, V., and Goncalves, J. 2013. A Network Science Approach to Modelling and Predicting Empathy. In Proceedings of the International Conference on Advances in Social Network Analysis and Mining, ACM, 1395-1400.

[32] Venkatanathan, J., Kostakos, V., Karapanos, E., and Goncalves, J. 2014. Online Disclosure of Personally Identifiable Information with Strangers: Effects of Public and Private Sharing. Interacting with Computers, 26(6), 614-626.

[33] Wölfer, R., Cortina, K.S., and Baumert, J. 2012. Embeddedness and empathy: How the social network shapes adolescents' social understanding. Journal of Adolescence, 35(5), 1295-1305.

[34] Zhuang, L., Jing, F., and Zhu, X.-Y. 2006. Movie Review Mining and Summarization. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, ACM, 43-50.