31005 ADVANCED DATA ANALYTICS

ASSIGNMENT 3 : TAKE HOME EXAM

RESPONSE TO QUESTION ONE

KATE MORAN 12403147

THE QUESTION

> "Following your graduation, you are hired by a polling organisation as a data analyst. As social media has exploded and transformed the way people interact with each other, it would be a great idea to use messages collected from social media to predict how the user can be converted to change his/her support. List three challenges to solving this problem. With reference to existing approaches, describe the design of your system. Discuss the ethical and social consequences of this study."

INTRODUCTION

Social media is used by the masses as a form of self-expression and communication and a tool for social participation. Its usage has resulted in large volumes of available data that, when aggregated and analysed, can be used to weigh public opinion. Researchers have attempted to apply this data usage to the political arena - to predict the popularity of candidates and the outcome of elections.

This response will look at the challenges presented when utilising social media data and past approaches to harnessing social media data for political prediction. It will also look at potential social and ethical consequences that this practice raises.

CHALLENGES

SOCIAL MEDIA USERS ARE NOT REPRESENTATIVE OF THE OVERALL VOTING POPULATION

The initial challenge posed by using social media data for political prediction is that social media users are not representative of any voting populations. Researchers would find it challenging to collect representative samples needed for political prediction. The difficulty is due to the age groups that use social media, the platforms they are present on, and how researchers collect their data.

Twitter is the data scientist's platform of choice for social media analysis, due to the ease with which researchers can collect data using their API (Avello, Metaxas & Mustafaraj 2011). However, Twitter only captures a small subset of the population (Wojcik & Hughes 2019). Many younger users prefer Snapchat or Instagram, and many older users choose to use no social media platforms. Therefore, researchers must carefully consider whether the predictions they form from social media data can be generalised to the entire voting population.

THE VOCAL MINORITY AND THE SILENT MAJORITY: SOCIAL MEDIA CONTENT IS NOT REPRESENTATIVE OF THE OVERALL OPINION OF THE VOTING POPULATION

Mustafaraj et al. (2011) identified that there is a spectrum of users who interact with social media in different ways. At the two ends of this spectrum, we have the vocal minority (those who share lots of content and often) and the silent majority (those who share once). They noted that there were significant differences between the two groups in the structure of the content they shared, as well as their objectivity. The challenge for pollsters is determining whether the sample they have taken is representative, not only of users but the voting population overall. Aggregated social media content may not be representative of social media users overall.

BEWARE BOTS AND PROPAGANDISTS

Bots are social media accounts that automatically interact with other users (Kollanyi, Howard & Woolley 2016). Political bots can swiftly disseminate messages on policy issues, political crises and elections. Alarmingly, they can pass as human users and are utilised by political actors to alter the debate. Identifying content that has been produced by a bot can be a challenge for researchers. The content generated by bots can muddy and pollute the observed data and influence analysis when attempting to make political predictions.

SYSTEM

DATA COLLECTION & PREPROCESSING

Twitter provides users with an API which is ideal for collecting data for this purpose. It allows users to extract tweets based on time, location and related words. Keywords for political prediction may include candidates' names or other hot topic policy issues and are generated by looking at term frequency. However, as Wang and Gan (2017) noted in their Twitter analysis, some keywords may not be sentimental or may take on a different meaning over time. It is essential to combine analysis with domain knowledge to select keywords.

SENTIMENT ANALYSIS

It is common across most methods to utilise sentiment analysis to calculate popularity. Sentiment analysis determines the tone behind words to gain an understanding of the opinions of the writer. There are a few types of sentiment analysis algorithms. Rule-based systems perform analysis based on a set of manually created rules and often use classical natural language processing techniques. These systems can be very naive and can become complex quite quickly. Alternatively, sentiment analysis can be conducted using machine learning techniques as a classification problem. The classifier is given text that has been labelled as positive, negative or neutral to train on (Monkey Learn 2019).

METHODS FOR PREDICTION

The most popular method cited in the research is Tumasjan's method (as cited in Wang and Gan 2017). The method calculates the popularity of a candidate using the following (Wang and Gan 2017):

$$popularity(a) = \frac{pos(a) + neg(b)}{pos(a) + neg(a) + pos(b) + neg(b)}$$

pos(a) and neg(a) are the number of positive and negative tweets for the (a) candidate. This method does not consider neutral tweets. Wang and Gan (2017) think that by ignoring neutral tweets, this may bias toward a candidate who is strongly supported by a small group of voters. Due to this, Wang and Gan present their modified formula for determining popularity:

$$popularity(a) = \left[\frac{pos(a)}{pos(a) + neg(a)}\right]\left[\frac{N(a)}{N(a) + N(b)}\right]$$

If there are more than two candidates, this formula requires scaling, so the sum of all popularities equals 100%.

ETHICS

When pollsters utilise social media data to predict political popularity, they should consider possible social and ethical issues and their consequences.

DATA COLLECTION

Pollsters should consider how they are sourcing and collecting their social media data to ensure consent is given and that they are not breaching the user's privacy. Users of social media platforms, especially Twitter, are posting in the public domain. Collecting and analysing this data could be considered to be consistent with the user's original expectations (Ballantyne 2018) because they were posting into a public forum. However, pollsters and researchers should not attempt to collect or utilise data in clandestine ways or from websites that do not allow this collection from their users.

POPULARITY PREDICTION TWEAKING

The process of predicting political popularity uses sentiment analysis and often involves machine learning techniques. As with most machine learning models, parameters can be tuned, and the model can be tweaked to change the model’s output. Therefore, it could be tempting for pollsters to use these as ‘levers’ to try to tweak their results. They may do this for many, such as to make their predictions fall into line with others (Bialik 2014). Even more sinister, they may also be aligned with a particular political candidate and might use their polls to manipulate public perception. These sorts of activities have the potential to undermine democratic elections, which has significent social consequences.

CONCLUSION

Using social media data to make political predictions has many challenges. Aggregated social media content is not necessarily representative of all social media users, and social media users are not representative of the voting population. Social media can also be the target of bots and spammers who can pass as humans to disseminate political messages which muddy and pollute the observed data. The process of using social media data to predict political popularity involves collecting the data using keywords and performing sentiment analysis. Classical NLP and machine learning techniques may be adopted to label social content as positive, negative or neutral. Several formulas exist to calculate popularity, with Tumasjan’s method being the most popular. The practice of utilising social media data also raises a number of ethical issues, namely issues surrounding privacy and consent as well as avoiding the manipulation of polls to maintain a healthy democracy.

REFERENCE LIST

Avello, D., Metaxas, P. & Mustafaraj, E. 2011, 'Limits of electoral predictions using twitter', Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, viewed 2 October 2019, <https://pdfs.semanticscholar.org/3ef4/3622a68e717228ce0698246e427339a4bbdc.pdf>.

Ballantyne, A. 2018, 'Where is the human in the data? A guide to ethical data use', GigaScience, vol. 7, viewed 5 October 2019, <https://academic-oup-com.ezproxy.lib.uts.edu.au/gigascience/article/7/7/giy076/5046607>.

Bialki, C. 2014, 'Pollsters say they follow ethical standards, but they aren't sure about their peers'. FiveThirtyEight, viewed 5 October 2019, <https://fivethirtyeight.com/features/pollsters-say-they-follow-ethical-standards-but-they-arent-so-sure-about-their-peers/>.

Kollanyi, B., Howard, P. & Woolley, S. 2016, 'Bots and automation over Twitter during the US election', Comprop Data Memo 2016.4, viewed 2 October 2019 <http://blogs.oii.ox.ac.uk/politicalbots/wp-content/uploads/sites/89/2016/11/Data-Memo-US-Election.pdf>.

Monkey Learn 2019, Sentiment analysis, viewed 3 October 2019, <https://monkeylearn.com/sentiment-analysis/>.

Mustafaraj, E., Finn, S., Whitlock, C. & Metaxas, P. 2011, 'Vocal minority versus silent majority: discovering the opinions of the long tail', 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust, pp. 103-110, viewed 3 October 2019, <http://ieeexplore.ieee.org.ezproxy.lib.uts.edu.au/stamp/stamp.jsp?tp=&arnumber=6113101&isnumber=6113084>.

Wang, L, & Gan, J. 2017, 'Prediction of the 2017 french election based on twitter data analysis', 2017 9th Computer Science and Electronic Engineering (CEEC), pp. 89-93, viewed 2 October 2019, <http://ieeexplore.ieee.org.ezproxy.lib.uts.edu.au/stamp/stamp.jsp?tp=&arnumber=8101605&isnumber=8101585>.

Wojcik, S. & Hughes, A. 2019, 'Sizing up twitter users', Pew Research Centre, viewed 3 October 2019, <https://www.pewinternet.org/2019/04/24/sizing-up-twitter-users/>.