Github:
https://github.com/Revesby/UTS_ML2019_Main/blob/master/A3_31005_12912616_Ali_Reezvy.pdf

Introduction

Given the growth of social media and it's increasing presence in household lifestyle, companies in the modern era of technology wish to extract deeper insights on consumer trends and how users shape opinions on a range of topics in the social media space. Questions such as how users form opinions and how support, or lack thereof, of these opinions can change are commonplace and different approaches have been taken from a data analytics perspective to predict these outcomes. In this report, we will discuss existing approaches that have been implemented in predicting how users can have their support converted and how these approaches can be combined into a system design. Additional ideas will also be discussed that could be further explored with theory, application and a proper budget allocated to innovation. Associated challenges will be discussed as well as potential ethical and social consequences that may arise as a result.

System Design: Sentiment Analysis

Sentiment analysis is a data mining process that identifies qualitative opinions expressed through different mediums of text, such as social media, and uses natural language processing to determine whether the user's attitude or opinion is positive, negative or neutral.

Based on experiments conducted by Agarwal et. Al, three types of models are used to conduct the sentiment analysis. The first model used in experimentation is the **unigram model**. The probability of an instance of a word is analysed by the unigram model from an independent sample. These results are then compared with data dictionaries to determine whether the text in question is classified as positive, negative or neutral.

The second model used in experimentation is the **feature-based model**. The first unigram model mentioned is used as a foundation whilst additional features are appended to the model. This is useful for classifying the 'sentiment' of the given text, which could be a tweet on the social media platform Twitter as an example. The additional features appended to the model could be specific hashtags that are associated with a given sentiment (#angry for negative sentiments) or the length of a Twitter thread which can indicate a user ranting about a certain issue.

The final model used in experimentation is a **tree-kernel based model**. A compact representation of texts from social media are created with numerous attributes and categories. A partial tree kernel is then used to compare and contrast trees by comparing with all possible sub-trees.

$$K(T_1, T_2) = \sum_{n_1 \in N_{T_1}} \sum_{n_2 \in N_{T_2}} \Delta(n_1, n_2),$$

where N_{T_1} and N_{T_2} are the sets of nodes in T_1 and T_2, respectively and $\Delta(n_1, n_2) = \Sigma_{i=1}^{|\mathcal{F}|} I_i(n_1) I_i(n_2)$, i.e. the number of common fragments rooted at the n_1 and n_2 nodes."

Implementing this system design around a social media platform to determine whether a user has converted their support is possible with given past data and previous classification of users into existing outcome groups. A classifier would be necessary to run on a given user's data from an arbitrary time period prior to the present time to predict accurately how support has changed and what factors are correlated with these changes in users.

Challenges: Communication Nuances

Sentiment analysis is ultimately an attempt to conquer the realm of qualitative analysis with a quantitative approach using data analytics, machine learning etc. With this assumption in mind, it is also worth noting that disparities in analysis occur which are difficult to measure in the quantitative environment leading to some challenges (that still remain unsolved in the modern day) in the analysis of social media opinions and what influences their change.

A frequent challenge in determining what causes a user to change their support is the use of sarcasm in social media. A typical sentiment analysis classifier will predict the user's sentiment to be positive in a given text wherein the user has posted a sarcastic display of support for a particular product/political opinion etc. Nuances of human communication are difficult to add as a feature in the feature-based model of sentiment analysis without explicit definitions, and hence this is a challenge present in predicting changes in support on user's platforms on social media.

Ethical and Social Consequences 300

Sentiment analysis and specifically natural language processing (NLP), a component of sentiment analysis, has ethical consequences that are still immature to fruitful discussions on solutions. Previously, NLP has never directly involved using humans and subjecting them to experimentation therefore the likelihood of a breach of ethics was low. However, Hovy et. Al argues that the increase in social media data and use in statistical modelling and machine learning opens up ethical issues of privacy and identification of individuals whose features are being measured in experiments to predict an outcome. '*NLP language – is a proxy of human behaviour… [people] can also be identified as members of specific groups by their use of subconscious traits…*' (Silverstein, 2003; Agha, 2005; Johannsen et al., 2015; Hovy and Johannsen, 2016).

On a social consequence standpoint, sentiment analysis suffers from topic exposure, which creates biases. If analysis consistently determined that language from a specific demographic was more difficult to process, it could underrepresent this group or create a classification where this group is conveyed as abnormal. Avenues for discrimination are very possible in this case.

References

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. 2011. 'Sentiment analysis of Twitter Data' [Online] Available at: http://delivery.acm.org/10.1145/2030000/2021114/p30agarwal.pdf?ip=14.200.88.132&id=2021114&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&__acm__=1570620242_35ac88d3c602418a fd45982fb7e6539b [Accessed 09 October 2019].

2. Mohey El-Din, D. 2016. 'A Survey on Sentiment Analysis Challenges'. Journal of King Saud University - Engineering Sciences. -. 10.1016/j.jksues.2016.04.002. Available at:

https://www.researchgate.net/publication/301649355_A_Survey_on_Sentiment_Analysis_Challenges [Accessed 09 October 2019].

3. '10 Sentiment Analysis Issues to Be Aware Of '[Online] Available at: https://www.adweek.com/digital/37705-sentiment-analysis/ [Accessed 09 October 2019].
4. Hovy, D., Spruit, S. 2016. 'The Social Impact of Natural Language Processing' [Online] Available at: https://www.aclweb.org/anthology/P16-2096.pdf [Accessed 09 October 2019].