Assignment 3:

Student Number: 11403868

Student Name: Nicholas Turnham

Course: 31005

Question:

"Marketing or advertising companies would be very interested in being able to predict whether a Twitter message will spread as a meme or not, and even better, construct it so that it will spread. Why is this a hard problem to solve? Describe two approaches using data analytics to predict whether a tweet will go viral or not. How would you validate these approaches? Discuss the ethical and social consequence of this study."

Assignment GitHub Repository: https://github.com/tunrham/UTS_ML2019_ID11403868/

Assignment Name: 'A3'

Introduction

The prevalence of technology in the role of communication and cultural connectivity has exploded in the 21st century through social networking platforms such as Facebook, Twitter, Reddit, and Instagram. These platforms have provided a unique medium for interaction, emphatically headlined by "Memes" – which are viral online media (either image, text, or video) spread across social networks (often featuring mutations). Due to the addictive and socially impactful nature of Memes, the ability to successfully manufacture them is desirable – especially in fields like Marketing and Politics.

Why are Memes difficult to predict?

Predicting the virality of Memes is an incredibly difficult task as it is difficult to manufacture and model. Difficulty in predictability relates to the complexities and intricacies involved in the popularization of Memes. Weng et al. (2013) suggest that to create a unique Meme it must: contain social and cultural relevance, gain traction amongst clustered groups of people, and follow a flow of osmotic diffusion between different segmented clusters. Successful Memes must also compete in a hyper-competitive, saturated, zero-sum market where an increasing social share of one Meme coincides with the diminishment of another.

Modelling the creation of virality is also difficult due to the black swan nature of a successful Meme. There is a monumental difference between current social media content, and the likelihood that any of that content will "go viral". Cheng et al (2014) outlines this difficult balancing act between predicting an incredibly unlikely outcome against artificially boosting the likelihood of said outcome in the training dataset – which would develop a model greatly diverging from real-world application.

Generalized Linear Model Methodology Predicting Tweet Virality

Jenders et all (2013) provides a solid baseline approach to predicting the virality of "Tweets" (the messaging system used Twitter), outlining both a methodology for predicting virality, and the attributes that are most important in manufacturing it. The dataset was drawn from both the user of the Tweet (including follower size, the list of accounts followed, and their previous Tweets) and the Tweet itself (including the sentiment of the message, the date of the message, and the number of reposts the Tweet has).

From here a Generalized Linear Model was used, measuring a scoring output for identifying the importance of different weights for Tweet virality (measured in reposts). Amongst these weights, number of initial followers (for the first poster) was deemed the greatest attribute for virality. Other attributes such as: the number of mentions the account and original post got, and the positivity of the sentiment of the post were also relevant factors in extreme virality.

Clustered Community Network Driven Random Forest Methodology Predicting Tweet Virality

Another method for predicting Meme virality was proposed by Weng et al (2014), which involved examining Memes and their attributes within a Clustered Community Network – likening Meme growth to that of infectious disease. These attributes were clustered by:

Network Features:
Focusing on the connectivity of Meme users, examining: the number of early adopters of a meme and the number of unaffected neighbours of early adopters.
Distance Features:
Focusing on theoretical distance between adopters of a Meme within a network; measuring the distance and coverage of a tweet within a community network.
Community Features:
Examining different communities, observing: the inter-community and intra-community infection and adoption rate of a Meme.
Growth Rate Features:
Focusing on the duration of time-steps a Meme exhibited during its growth.

After establishing the most important attributes (and sub-attributes) for Meme virality, a Random Forest was used, creating 300 trees using five sub-attributes from the attribute feature classes. This was then tested against five other baseline models of virality prediction - identifying both virality and lack thereof through accurately forecasting Meme Usage and Meme Adoption. The model was more robust than baseline models, avoiding skewing by adjusting for imbalanced class sizes.

Validating Approaches to Meme Virality Prediction

The processes for observing Meme virality are difficult to accurately forecast and validate, due to the attributes of virality constantly evolving through a state of flux. Jenders et al (2013) posit that their methodology should only be used as a starting point because of the constantly evolving and intertwined nature of virality deriving attributes. This is a common problem present throughout virality forecasting – many methods only observe a start and end point – instead of observing virality at each stage of a Meme's life cycle.

Cheng et al (2014) examine virality chains (known as "cascades") in a sequential life cycle manner, using binary classification at each time-step. Whilst this study examined Facebook post virality instead of Tweets, it is a methodologically applicable study that considers how changing input variables effect a Meme over its life cycle. Given the difficulty to validate virality forecasting due to the liquidity and complexity of input variables, methodology that includes time-step information represents a fundamentally improved baseline for future research.

Ethical Considerations of Manufacturing Memes for Marketing

Marketing applications of Meme virality border numerous ethical quandaries and paradigms. On one hand, a cleverly created viral campaign can qualify as targeted interactive media-based marketing. Using a "Consequentialism" based ethical framework (assessing actions based on the results of the action), it can be argued that furthering research for Meme virality has such a minimalistic consequential societal reaction that we should be indifferent towards any potential field research.

Conversely, using a Kantian "Rights Approach" analysis of Meme virality outlines reasons to avoid further research. This approach assumes people deserve the right to think and decide freely. Darmoc (2018) explores how attempting to create virality impedes on this right – as marketers are attempting to create addictive content, crossing an ethical boundary. Given the multiple applicable paradigms, it is difficult to quantify the ethicality of Meme virality – based on which ethical framework the question is posed from.

Conclusion

Given the popularity and prevalence of social media platforms, creating content that has high virality represents a desirable research goal for multiple parties. Models such as Generalized Linear Modelling and Random Forests (using a Clustered Community Network for tree attribute identification) represent fundamentally sound, yet stagnant predictive methodology – failing to account for the dynamic nature of Meme life cycle. For evaluation purposes, analysing multiple points of the virality life cycle (such as with Cascade Binary Classification by time-step) offers a more robust methodology for validating Meme diffusion. Further study of Meme virality however, should be considered using the framework that the industry and its players choose to evaluate ethical decision making under – as there are varying paradigms.

References

Weng, L., Menczer, F., Ahn, Yang-Yeol., 2013, *'Virality Prediction and Community Structure in Social Networks*', Sci. Rep 2013, sourced https://www.nature.com/articles/srep02522.pdf at 5/10/2019

Cheng, J., Adamic, L., Dow, P., Kleinberg, J., Leskovec, J., 2014, '*Can Cascades be Predicted*?', Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, ACM, sourced http://www.cs.cornell.edu/home/kleinber/www14-cascades.pdf at 5/10/2019

Jenders, M., Kasneci, G., Naumann, F., 2013, '*Analyzing and Predicting Viral Tweets*', Proceedings of the 22nd International Conferend on World Wide Web, sourced https://www.researchgate.net/publication/262166912_Analyzing_and_predicting_viral_tweets at 5/10/2019

Darmoc, R., 2018, '*Marketing Addiction: The Dark Side of Gaming and Social* Media', Journal of Psychological Nursing and Mental Health Services 2018, sourced https://www.healio.com/psychiatry/journals/jpn/2018-4-56-4/%7Bb1930695-c1a4-45f4-a352-8b4c4945b104%7D/marketing-addiction-the-dark-side-of-gaming-and-social-media.pdf at 5/10/2019