Advanced Data Analytics

Assignment 3 – Take Home Examination

Chosen Problem – Question 3

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 *social and ethical consequences* of the study.

Predicting Tweet Virality – Why This is a Hard Problem to Solve

Twitter, being a social network platform, is fundamentally subject to the enormous complexities of social psychology, as well as those of culture and current events. Thus, what might be initially seen as the simple task of solving for the virality of a tweet turns out to be deceivably challenging, as it seems that any highly effective solution must, in some way, strongly take into consideration those social and cultural complexities that are associated with human interaction. Although it would be possible to construct a model that only utilises what might be considered "obvious" features (such as the number of followers a user has, tweet length, number of mentions, etc)[4] for the purpose of having at least some capacity for predicting tweet virality, this insight would be ineffective in facilitating a method with which marketing and advertising companies would be able to meticulously design highly effective, diverse and specific tweets that consistently go viral. The challenge of this problem can therefore be seen to exist on a spectrum, depending on how sophisticated one wishes the solution to be. The more sophisticated the solution, the more effectively advertisers and marketers can utilise it towards their own ends.

Given that this problem is being tackled within the context of data analytics, it is crucial to be able to arrive at these insights using data that is in quantifiable form. As the complexity of the solution increases, so too does the complexity of psychological and cultural considerations, and thus the task of expressing the phenomena strictly in terms of quantifiable data becomes increasingly difficult. In order to discern which approaches are likely to be most capable of facilitating these more sophisticated solutions, a careful investigation regarding the general framework, strengths and weaknesses of these approaches is necessary. By identifying which approach most closely caters to the nature of the problem, we create a foundation from which a realistically implementable solution is possible. The following sections detail two possible approaches, which are each validated in terms of the problem.

Data Analytics Approach 1 + Validation – Generalised Linear Model

We suggest that the generalised linear model is a possible candidate for enabling sophisticated solutions to the problem of predicting tweet virality, but only to the extent at which nuanced features can be quantifiably expressed. The task of quantifiably transcribing those relevant insights in social psychology and cultural affairs may not be immediately obvious, however successful progress has already been made in this regard by [4]. Specifically, the notion of sentimentality contained within a tweet was investigated in terms of how it might correlate with retweet frequency (virality). In order to quantifiably discern the sentimentality of a tweet, an algorithm was created that associated the meanings of words in English dictionaries and thesauruses with numerical values between the range of –5 and +5, denoting negative and positive sentimentality respectively. This clever solution enabled the authors of the paper to investigate, within a data analytics context, a seemingly subjective and unquantifiable phenomena, and discover that the sentimentality of a tweet did in fact strongly correlate with retweet frequency. Specifically, it was discovered that tweets with a negative sentimentality, on average, accrued more retweets. If one can similarly devise methods of quantifying other relevant factors which, at first, appear to be completely subjective in nature, then it would be possible, overtime, and in conjunction with those more "obvious" aforementioned features, to be able to reliably predict the virality of a tweet based on how they relate in terms of a generalised linear model. If a tweet's virality is simply defined as whether a particular tweet has accrued a number of retweets which exceeds some predetermined threshold, then a highly effective and realistically implementable solution to the problem is possible through a regression-based approach, specifically generalised linear regression (as opposed to alternatives such as simple linear regression) due to its increased capacity to handle categorical features.

Data Analytics Approach 2 + Validation – Bayes Model

An alternative to the generalised linear model that also seems capable of handling more sophisticated types of data is the Bayes Model, as demonstrated by [3], who demonstrated, with a deliberate consideration of a subtle fact regarding humans, namely that different people are interested in and thus tend to gravitate towards different things/topics, was still able to develop a model that demonstrated high levels of prediction accuracy. By considering the interests of twitter users (which is publicly available) in conjunction with the key words in tweets that relate to topics that are like those interests, a quantifiably expressible positive match can be identified with which a Bayes model is well suited for the task of combining the data into a theoretical framework. This makes it suitable for eventuating generalised solutions to the problem of virality prediction, making it even more attractive to advertisers and marketers, since the insight may be applied to more than one domain. Another advantage that the Bayes model provides is its capacity to determine conditional inferences and probabilities between data features, as this could potentially provide a means for discovering complementary features that might also be later used a convenient tool for designing highly effective tweets.

Social and Ethical Consequences

There has already been much discussion regarding the ethics of using data analytics technologies, especially those of machine learning, for advertising, marketing and other strictly commercial purposes [8]. The fundamental consideration surrounding this discussion is primarily regarding individual privacy, and to what extent the use of machine learning would begin to border on exploitative. Many people deeply value individual freedom, which is something closely linked to individual privacy. Some people argue that the use of machine learning has the potential to covertly cause breaches in individual privacy, and thus violate individual freedom. Many organisations use machine learning techniques in order to study customer data for the ultimate purpose of maximising profits. This has already been applied in many areas, most notably in that of recommender systems, such as that found on websites such as YouTube. If appropriate legal actions are not made regarding the limits of what machine learning may be used for, as well as in what ways data may be collected, then organisations will likely continue to expand its usage into other areas that include social media virality prediction. Specially, regarding Twitter retweet prediction, it is not difficult to imagine that organisations may use machine learning techniques, such as those explained in the previous sections, for the purpose of creating a method that allows for the consistent output of viral tweets, ultimately leading to an increase in the organisation's noticeability and therefore likely profits as well. Given that it is theoretically possible to utilise sophisticated facts about human nature that come from fields such as social psychology and cultural history, it is arguable that there ought to be some extent from which further inquiry for the mere purpose of maximising profits might be considered unethical due to the obvious potential for deep privacy breach. What constitutes an ethical use of machine learning for these purposes, as opposed to an unethical use, is currently up for debate. In fact, many aspects and applications of artificial intelligence comprise core ethical problems that are being disputed to this day [8].

References

[1]EliteDataScience. (2019). *Modern Machine Learning Algorithms: Strengths and Weaknesses*. [online] Available at: https://elitedatascience.com/machine-learning-algorithms [Accessed 7 Oct. 2019].

[2]Hong, B. (2019). *Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network - IEEE Conference Publication*. [online] Ieeexplore.ieee.org. Available at: https://ieeexplore.ieee.org/abstract/document/5590452/ [Accessed 7 Oct. 2019].

[3]Huang, D. (2019). *Analyzing User Retweet Behavior on Twitter*. [online] Available at: https://www.researchgate.net/publication/282925762_Retweet_Behavior_Prediction_in_Twitter [Accessed 7 Oct. 2019].

[4]Jenders, M. (2019). *Analyzing and predicting viral tweets*. [online] Hpi.de. Available at: https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/publications/2013/Analyzing_and_Predicting_Viral_Tweets.pdf [Accessed 7 Oct. 2019].

[5]Kowalczyk, D. (2019). *Scalable Privacy-Compliant Virality Prediction on Twitter?*. [online] Ceur-ws.org. Available at: http://ceur-ws.org/Vol-2328/1_paper_27.pdf [Accessed 7 Oct. 2019].

[6]Lexalytics. (2019). *Machine Learning for Natural Language Processing - Lexalytics*. [online] Available at: https://www.lexalytics.com/lexablog/machine-learning-vs-natural-language-processing-part-1 [Accessed 7 Oct. 2019].

[7]MacCarthy, M. (2019). *How to address new privacy issues raised by artificial intelligence and machine learning*. [online] Brookings. Available at: https://www.brookings.edu/blog/techtank/2019/04/01/how-to-address-new-privacy-issues-raised-by-artificial-intelligence-and-machine-learning/ [Accessed 7 Oct. 2019].

[8]Privacy, A., Videos, P., Services, O., Assessments, P., Reviews, P., Workshops, P., Workshops, I., Training, O., Certifications, I., topics, T., Kits, C., topics, R., Handbook, F., Are, W., Johnston, A., Casley, M., Calleia, A., Wilson, S. and Certifications, I. (2019). *The ethics of artificial intelligence: start with the law*. [online] Salingerprivacy.com.au. Available at: https://www.salingerprivacy.com.au/2019/04/27/ai-ethics/ [Accessed 7 Oct. 2019].

[9]Support.sas.com. (2019). *SAS/STAT(R) 9.2 User's Guide, Second Edition*. [online] Available at: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_introbayes_sect006.htm [Accessed 7 Oct. 2019].