Using Machine Learning to Predict whether a Tweet will go Viral for Marketing and Advertising Purposes

By Poojitha Kalva - 12572350

Introduction

An advancing component of data analytics is companies' ability to use machine learning in social media to their benefit in marketing and advertising. To achieve this, studies surrounding how Twitter messages, tweets, spread as viral messages is imperative. In studying this, marketing and advertising companies may be able to identify the factors contributing to the viral nature of these tweets and consequentially, construct tweets in a manner that ensures they do spread.
In this paper, we delve into analytical research and experiments conducted to test the capability of machine learning to effectively and accurately predict the viral nature of tweets and in turn, discuss the benefits and downfalls of each approach.

The Problem

The main challenge in solving the problem of predicting whether a Twitter message will spread as a meme or not is the numerous aspects and perspectives of a tweet to consider. We need to effectively identify the multi-faceted nature of tweets including the language it is constructed in, the use of features such as hashtags and mentions, the user and their attributes and the aspect of retweets. To efficiently cater to these aspects, we examine all influential factors in viral tweets below.

Influential Factors in Predicting Viral Tweets

In analysing the viral nature of tweets, various research proves that there are numerous features of the construction of the message that contribute its rate of spread – in particular, this can be categorised as structure, content and sentiment, as well as distinguishing them as user features and tweet features. It is vital to understand these influential factors before diving into the specific machine learning approaches that serve as prediction models.

Structural analysis focuses upon the structure of the social network and specifically, identifying features relevant to influential users – this includes, number of followers, retweets and mentions (Jenders, Kasneci and Naumann (2013)). In analysing these attributes of a tweet, or more so the features of the user, we can also define these as user features, similar to Bunyamin and Tunys (2016). From another perspective, analysing viral tweets also involves investigating the content of the message. Thus, content analysis revolves around the components of the tweet including but not limited to the use of hashtags, hyperlinks, and username mentions. Sentiment analysis is yet another category of tweet features that influences the viral nature of tweets. It analyses the emotions expressed through a message and these sentiments can be classified as positive, negative and neutral (Jenders et al. (2013). As these two categories analyse solely the features of the tweet itself, we are able to classify these as tweet features (Bunyamin and Tunys (2016)).

Predictive Models

1. Random Forest

Another machine learning approach for classification that can has been investigated for predicting whether a tweet will go viral is Random Forest. This classifier is essentially a combination of numerous decision trees serving as an ensemble to increase accuracy of performance and prediction. To explain the logic behind multiple trees, each decision tree performs its own function of classification and the Random Forest algorithm collates these classifications and identifies the most voted for prediction. In a way, this reduces the chance of error and misclassification if there were not numerous trees to cross-check – the weak classifiers merge together to form a stronger classifier.

To understand the usage of Random Forest in predicting viral tweets, we explore previous research by Weng, Menczer and Ahn (2014) and Bunyamin and Tunys (2016). When considering the Random Forest model, it is imperative to begin with identifying relevant features to use in building the predictive model – this could be done in various ways as mentioned in the above section. In the case of Weng et al. (2014), they employ a different method of categorising features – Basic Network Features, Community Features and Growth Rate Features. On the other hand, Bunyamin and Tunys (2016) distinguish between features as user and tweet features and detail how the selection of these two feature types affects the prediction results. After examining both approaches, it seems more reliable to implement both user and tweet features in the Random Forest Model – it is evident that Weng et al. do not sufficiently examine the content of tweets or consider these features in the selection of influential factors but rather focus solely on network structure. In doing more research, including the work Naïve Bayes work by Jenders et al., tweet features prove to be imperative in studying viral nature and thus, we can consider this a downfall in the work of Weng et al. and a positive in Bunyamin and Tunys' studies.

In implementing the algorithm, the count of retweets has been used as a measure of a tweet's popularity and viral nature. Bunyamin et al. (2016) utilise a 5 fold cross-validation in their method.

2. Naïve Bayes

Used for classification and predictive modelling, the Naive Bayes algorithm is derived from Bayes Theorem – it is essentially a probabilistic model that assumes strong independence between attributes. The Naive Bayes classifiers are most effectively used with text classification problems and are known to be highly scalable in real-time prediction and thus, beneficial in the context of social media. In assuming independence between features, the model supposes that changing the value of one attribute will not influence other attributes.

Before exploring the model itself, we can identify the features that were chosen to create the predictive model. In the case of Jenders, Kasneci and Naumann (2013), we notice that they have utilised both, user features and tweet features – specifically, they have incorporated all three categories of content, structure and sentiment analysis. These features are sentiment valence,

categorised as positive/negative, tweet character length, number of mentions, hashtags and followers, emotional divergence and number of URLs (Jenders et al. (2013)).

In this model, as a method to separate viral and non-viral tweets, a threshold of *T* was used by Jenders et al. (2013) and virality was measured by the number of retweets. The method also employs a tenfold cross-validation.

Validation and Evaluation of Prediction Models

Validation

To ensure validation, both models use the k-fold cross validation method. This enables us to judge the model's performance on a larger dataset using a limited set of data. Essentially, we split the dataset into *k* number of groups, select one group as the test data and use the remaining as training data, apply the model to the training set and test on test data and evaluate the performance. This process is repeating on the remaining *k* groups. *K* is the only parameter in this validation technique and simple refers to the number of iterations. In the case of Naïve Bayes Model, a 10 Fold Cross-Validation technique has been performed and a 5 Fold Cross-Validation Technique has been used to validate the Random Forest Model.

Evaluation

Both predictive models can be evaluated through their F1 scores which is essentially a measure of accuracy which considers both precision and recall. Precision is a measure of true positives amongst all classified positives (the summation of true and false positives) and Recall is the proportion of true positives against all positives (the summation of true positives and false negatives) (Powers, D.M.W., 2011).

The results of the Random Forest Model can be evaluated by using the F1 Score. In measuring accuracy, Bunyamin and Tunys' model resulting in a F1 Score of 74.47. For the Naïve Bayes Model, the F1 Score results in 0.916.

Ethical and Social Considerations

A very common ethical concern in machine learning has always been privacy – this is an especially increased problem when we apply data analytics to social media for the purpose of marketing and advertising. It can be highly risky for businesses to have the capability to identify individual components common within viral tweets and use these to their benefit in constructing a tweet or meme that is known to have a high likelihood of going viral. In a way, this could be seen as businesses forcing social media users to view their company in a certain light rather than allowing them to form their own opinion. However, it must be noted that all data used in this research was ethically obtained through Twitter API and is public information, so the concern lies more so in how marketing and advertising companies use the data.

References

Weng, L., Menczer, F. and Ahn, Y.Y., 2013. Virality prediction and community structure in social networks. Scientific reports

Bunyamin, H. and Tunys, T., 2016. A comparison of retweet prediction approaches: the superiority of Random Forest learning method. Telkonika (Telecommun Comput Electron Control)

Weng, L., Menczer, F. and Ahn, Y.Y., 2014, May. Predicting successful memes using network and community structure. In Eighth international AAAI conference on weblogs and social media.

Petrovic, S., Osborne, M. and Lavrenko, V., 2011, July. Rt to win! predicting message propagation in twitter. In Fifth International AAAI Conference on Weblogs and Social Media.

Jansen, B.J., Zhang, M., Sobel, K. and Chowdury, A., 2009. Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology, 60(11)

Acharya, A., Beaty, K.A., Jain, P. and Manweiler, J.G., International Business Machines Corp, 2017. Method for real-time viral event prediction from social data. U.S. Patent

Jain, P., Manweiler, J., Acharya, A. and Choudhury, R.R., 2014, May. Scalable social analytics for live viral event prediction. In Eighth International AAAI Conference on Weblogs and Social Media.