ISSN: 2581-3404 (Online)
International Journal of Innovative Research in Technology and Management, Vol-4, Issue-4, 2020.

A Survey on Online News Popularity Prediction Using Social Media

Smriti Shubham[1], Ritesh Kumar Yadav[2], Varsha Namdeo[3]

Department of Computer Science & Engineering[1,2,3]

SRK University, Bhopal, (M.P.), India[1,2,3]

ABSTRACT

Social media comprises interactive applications and platforms for creating, sharing, and exchange of user-generated contents. The past ten years have brought huge growth in social media, especially online social networking services, and it is changing our ways to organize and communicate. It aggregates opinions and feelings of diverse groups of people at a low cost. Mining the attributes and contents of social media allows us to discover social structure characteristics, analyze action patterns qualitatively and quantitatively, and sometimes the ability to predict future human-related events. In this paper, we firstly discuss the realms which can be predicted with current social media, then overview available predictors and techniques of prediction, and finally discuss challenges and possible future directions. Our work can help online news companies to predict news popularity before publication.

Keywords: Online News, Popularity Prediction, Social Media, Machine Learning.

Introduction

With the rapid growth of online news services and social media, it is very beneficial if we could determine readers' unseen behavioral patterns. It is also useful to shed light on readers' intentions and to predict the popularity of the online news, which means, whether, the online news will receive a significant amount of reader's attention. It is important to give pre-information to the media workers (authors, advertisers, etc.) to deal with each article according to its popularity without any influence from.

For this paper, we intend to make use of a large and recently collected dataset with over 39000 articles from the Mashable website [1]. For research, various machine- learning algorithms and other feature selection techniques were applied to first select informative features and then analyze and compare the performance of several machines-learning algorithms.

Machine learning algorithms can play an important role to process, analyze, understand, and predict the best choice for the task depending on some or all features. Various methods are available for feature selection such as Genetic Search, Best First, Greedy Stepwise, and others that can be applied using the Weka tool. The most famous machine learning algorithms [2] are Decision Tree, Random Forest Classification, Naive Bayes, K Nearest Neighbor (KNN), Bagging, Logistics Regression, Support Vector Machine (SVM), Artificial Neural Network (ANN) or Multilayer Perception, etc. This paper uses most of these machines learning an algorithm utilizes Precision, Recall, and F-measure as an evaluation measurement for results to find the best one. Results are compared with the previous work on the same dataset.

Social media are platforms that allow common persons to create and publish content. Two worldwide popular social media websites, Twitter and Facebook, demonstrate its explosive growth and profound influence. Both Twitter and Facebook are in the top 10 most-visited websites in the world according to Alexa ranking [3].

Facebook has more than 800 million active users [4], and by March 2011, on Twitter, there were about 140 million information pieces created and transferred daily [4]. Other specialized social media are focused on entertainment, sports, finance, and politics.

Since many users are sharing their opinions and experiences via social media, there is an aggregation of personal wisdom and different viewpoints. Such aggregation has limitations as viewpoints are subject to change with time. In a sense, the social media prediction problem is paralleled by the prediction of financial time series based on history, which has its uses in trading. In general, if extracted and analyzed properly, the data on social media can lead to useful predictions of certain human-related events. Such prediction has great benefits in many realms, such as finance, product marketing, and politics, which have attracted an increasing number of researchers to this subject. The study of social media also provides insights into social dynamics and public health. A survey provides us perspective and helps carry out further research.

II Related Work

Predicting and evaluating the popularity of online news have been studied extensively in numerous papers.

Ren et al. [6] applied many machine learning algorithms using Online News Popularity Dataset [1] from the UCI machine learning repository. They used Mutual information and Fisher criterion for feature ranking and selected 20 features that got the best rank.

Many machine learning algorithms were used in [6] namely: Linear Regression, Logistic Regression, Support Vector Machine, Random Forests, K-Nearest Neighbors, SVR (Linear Kernel), REPTree, Kernel Partial Least Square, Kernel Perceptron, and C4.5 Algorithm. They applied with 5-fold cross-validation. As a classification model, logistic regression achieved a decent accuracy, better than most of the models. Random Forest had the best result for this classification problem. They were able to utilize a different number of decision trees and a different number of features for each decision point. Therefore, they got an accuracy of about 66% in Linear Regression and an accuracy of about 55% in SVM with a polynomial kernel with d = 9. Finally, they changed the number of trees continuously from 5 to 500. The accuracy reached a limit of 69%, which is the best among all algorithms.

Frenandes et al. [7] used an Online News Popularity dataset in implementing various machine-learning algorithms namely: Random Forests, AdaBoost, Support Vector Machine, K-Nearest Neighbor, and Naïve Bayes. AdaBoost and SVM obtained an accuracy of about 66%. However, KNN, Naïve Bayes obtained an accuracy of about 62%, and the best result was achieved by a Random Forest (RF) i.e. with an overall area under the Receiver Operating Characteristic (ROC) curve of 73%, which corresponded to acceptable discrimination and an accuracy of about 67%.

In [8] introduced the work which predicts the volume of remarks of news stories which is before distribution by utilizing 5 capabilities. Forecast task is tended to as a 2 phase grouping task. 1) Binary grouping distinguishes articles that can get remarks. The grouping is performed for 2 classes for example with remarks versus without remarks. 2) The binary arrangement gets the yield from the initial step to mark the articles as "low" or "high" remark volume. Its outcome shows great execution for the past assignment yet execution debases for the last undertaking. The information comprises the substance from 7 online newsagents. The substance changed over into an amassed structure. The dataset comprises of 290375 articles, and 1894925 remarks. 5 gatherings of highlights are viewed as which are surface, total, printed, semantic, and genuine world. For each wellspring of a database, the creators made a preparation set and testing set. The preparation set comprises of articles distributed from Nov 2008 to Feb 2009 and the test set comprises of articles distributed in

March 2009. Irregular Forest calculation is utilized. F1 score and level of accurately characterized occurrences are assessed. The outcome is estimated with the Kappa measurement.

In [9] utilized computational phonetics to foresee consequently the effect of the news on open observation for political competitors. The framework utilized everyday paper articles to foresee changes in popular sentiment. Different sorts of highlights are intended for this issue. Barely any techniques for news include extraction are investigated with the market history benchmark. The news framework improves the expectation of the market over gauge advertise frameworks. The framework works iteratively. Every day, the news is utilized to build another occasion. A strategic relapse classifier is utilized to prepare for every earlier day and the subsequent classifier anticipated the value development of the new occurrence for example benefits or misfortunes cash. The outcome shows that news stories can be mined to foresee changes in the assessment of general society.

In [10] tended to the issue of prevalence expectation of online recordings partook in web-based social networking. We demonstrate this difficult assignment can be moved toward utilizing an as of late proposed profound neural system models. We cast the fame expectation issue as a characterization errand and we plan to unravel it utilizing just viewable signs removed from recordings. With that in mind, they recommend another strategy dependent on an LRCN that consolidates the sequentiality of the data in the mock-up. Results got on a dataset of more than 37,000 recordings distributed on Facebook show that utilizing our strategy prompts over 30% improvement in expectation execution over the conventional shallow methodologies and can give significant bits of knowledge to content makers.

In. [11] rewarded the ubiquity of online recordings as time arrangement over the given time frames and propose a novel time arrangement model for prevalence forecast. The proposed model depends on the relationship among's ahead of schedule and future prominence arrangement. Exploratory outcomes on genuine information have shown that the proposed model outflanks a few existing fame forecast models.

In [12] foresee the prominence of online substance dependent on highlights that can be seen by an outside client, including the number of remarks and the quantity of connections in the primary hours after the substance distribution. This work can anticipate lifetime dependent on a few strings (5–6 days) and a few client remarks (2–3 days). It is an enhanced paper for [10] utilizing endurance examination.

In [13] utilized a hereditary calculation to get the ideal characteristics and further arranged the information utilizing various classifiers and got the most elevated precision of 91.96% with NaiveBayes classifier.

III Prediction methods

In this section, we discuss some methods used in prediction with social media.

Regression method: Regression methods analyze the relationship between the dependent variable, prediction result, and one or more independent variables, such as the social network characteristics. The regression model could be linear and non-linear. But the linear model seems to describe the relation best [14]. Thus most times, we use the linear regression models, rather than non-linear ones, such as exponential, logarithmic, and polynomial models. In a linear regression model, the variables could be raw or transformed data. For example, between the early and late popularities of posts on digg.com, the correlation is based upon the logarithmically transformed data [15]. Besides, sentiment data does not work well in the regression models for movies. Currently, this is the simplest and most used method.

Bayes classifier: Bayes classifier is a probabilistic classifier using Bayes' theorem. Based upon the priori probability of the prediction event, Bayes

classifier uses the Bayesian formula to calculate its posterior probability, that the object belongs to the result classes, and then select the class with the largest posterior probability, as the event is most likely to have that result. If the prediction result is discrete, the Bayes classifier can be applied directly. Otherwise, the prediction result must be discretized first. This classifier has an assumption that the predictors must be conditionally independent. There is no solid evidence always that the discussed metrics satisfy this assumption.

K-nearest neighbor classifier: K-nearest neighbor classifier, one of the simplest machine learning algorithms, tries to cluster the objects according to their distance to others. Commonly we use the Euclidean distance and Manhattan distance. For two entities p = $\{p1, p2, \dots, ps\}$ and q = $\{q1, q2, \dots, qs\}$ with n-dimensional feature vector, the Euclidean distance is computed as:

Artificial Neural network: Artificial neural network is a computational model to simulate the human brain. An artificial neural network consists of lots of artificial neurons. And these neurons could belong to many interconnected groups, including the input layer, hidden layer, and output layer. The input layer is responsible for receiving raw data and transmitting them to the next layer. The output layer will give us the final prediction result. Thus using an artificial neural network to do prediction, our major task is choosing the network structure and designing the hidden layer. In addition to using them to predict directly, the Self Organizing Map (SOM), one kind of artificial neural network, could be used for features' dimensionality reduction for further analysis.

Decision tree: Decision tree is a visual technique in data mining and machine learning. Traveling from a root node to leaf, one entity will get the prediction result. The classification tree and regression tree are two basic and major types of decision trees. Classification tree analysis is applied when the prediction output is discrete classes. And the regression tree is used when a predicted outcome is a continuous value. Unlike the artificial neural network being a black-box model, the decision tree is a white box model, which could be relatively easily explained. Besides, the decision tree works well with dummy variables and empty variables.

Model-based prediction: This possibly is the hardest way to do prediction. We have to build a mathematical model on the object before prediction, which requires deep insight into the object. At this point, we do not know enough about social media to develop effective models for them. Even though there is some progress in modeling, model-based prediction remains an open and challenging topic.

IV Research Scope

As an emerging research topic, prediction with social media faces many challenges. Here we point out some urgent and important future works.

Using sociological theory to interpret predictors: Currently, researchers choose predictors using the trial and error method. We know neither why these predictors are better than others, nor how these predictors could predict the result. Not knowing the background logic between these metrics and the final prediction result, we just use a collection of metrics to be trained on test data, find out which ones have the highest coefficients, and use them to compose the prediction model. Consequently, lacking a solid supporting theory, we cannot be sure that one model, which works well in one case, could be applied to other situations with the same accuracy. That's why some models show good performance in one election prediction, but completely fails in another one. To guarantee our model has good performance in all cases, we need to know the logic and theory behind the model.

Trying more prediction methods, most researchers use simple methods such as linear regression analysis. These methods are known to work well under some conditions. Social media is produced on a complex system and thus more likely than not the predictors and prediction outcomes have a non-linear correlation. Furthermore, combination of methods might lead to breakthrough. In such combination, a surface learning agent, such as

instantaneously trained neural networks, quickly adapts to new modes and emerging trends on social media. And a deep learning agent focuses on long-term patterns. In a nutshell, we should try some non-linear methods and find out the suitable methods and/or combinations for each prediction realms.

Modeling on predictions with social media: We are far from knowing everything about social media. For instance, there are different kinds of prediction objects which show different features. Taking recommendation adoption as an example, the recommendation on DVDs is more likely to be accepted than that on books. But there is still no universally accepted conclusion about why these differences exist. This lack of understanding adds to the difficulties of modeling. Formal modeling could be necessary and helpful to understand and investigate the features and behaviors of prediction techniques.

Semantic analysis system for social media: Although semantic analysis is not a necessary part of the prediction methods, it is frequently used. Thus the accuracy of semantic analysis is critical to the prediction performance. The semantic analysis could be based upon lexicon or previous statistics. In terms of lexicon, compared with the natural and formal English language, social media content has a similar structure, but many different words, such as “lol”, which short for “laughing out loud”. This Internet slang affects the semantic analysis system because the lexicon in most existing systems is designed for well-written English. Besides, Internet slang evolves quickly and chaotically. The SOM, which sometimes is used to construct thesaurus as an unsupervised or semi-supervised clustering method, could be helpful in this issue. These methods firstly label some posts manually and then use statistical models, such as naïve Bayes classifier, to mark other posts according to statistical features of labeled ones. In some forms of social media, such as micro-blogging, the length of a post is so short that it shows no significant statistical characteristics.

IV Conclusions

What makes online news well known is as yet an exploration subject, however scarcely any focuses are found. The news content, for example, news, news, or quality, assumes a significant job in the prevalence of news achievement. There are other significant factors, for example, news, exposure, social effect, and so on. Variables influencing message prominence are significant for making increasingly precise prescient models. Social sharing is one of the extra impetuses of client consideration. Social associations inside a webpage assume a significant job in the ubiquity of online news.

In this paper, we presented a survey of prediction using social media. We also gave an overview of prediction factors and methods and listed challenging problems and areas for further research. Although prediction using social media is only an emerging research topic and its results have relatively low accuracy, it has created a new way for us to collect, extract and utilize the wisdom of crowds in an objective manner with low cost and high efficiency.

References

[1] Anon, 2016. UCI Machine Learning Repository: Online News Popularity Data Set. Archive.ics.uci.edu.

[2] Crisci, C., Ghattas, B. and Perera, G., 2012. A review of supervised machine learning algorithms and their applications to ecological data. Ecological Modelling, 240, pp.113-122.

[3] Alexa Internet Inc, "Alexa Top 500 Global Sites". http://www.alexa.com/topsites. [Accessed July 4, 2020].

[4] Facebook, "Statistics". http://www.facebook.com/press/info.php?statistic, [Accessed July 4, 2020]

[5] Twitter, "#numbers". http://blog.twitter.com/2011/03/numbers.html. [Accessed July 4, 2020]

[6] Ren, H., and Yang, Q., 2012. Predicting and Evaluating the Popularity of Online News.

[7] Fernandes, K., Vinagre, P., and Cortez, P., 2015. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Progress in Artificial Intelligence, pp.535-546.

[8.] Lerman, K., Hogg, T., "Using a model of social dynamics to predict popularity of news", In Proc. of WWW, ACM (2010), pp. 621-630, April 2010.

[9.] Lee Jong Gun, Sue Moon, and Kave Salamatian. "An approach to model and predict the popularity of online contents with explanatory factors." Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE, 2010.

[10.] Yangjie Yao, Aixin Sun, "Are Most-viewed News Articles Most-shared?", pp. 1-12, 20 June 2013.

[11.] UCI Machine Learning Database, archive.ics.uci.edu/ml/datasets/Online+News+Popularity, May 2015.

[12.] Manos Tsagkias, Wouter Weerkamp, Maarten de Rijke, "Predicting the Volume of Comments on Online News Stories", Proceedings of the 18th ACM Conference on Information and Knowledge Management, ACM, Hong Kong, China, pp. 1765-1768, 2009.

[13.] Tomasz Trzciskil, Pawel Andruszkiewicz, Tomasz Boche´nski and Przemyslaw Rokita, "Recurrent Neural Networks for Online Video Popularity Prediction", Springer, 2017.

[14] L. Liviu, "Predicting Product Performance with Social Media," Informatics in education, vol. 15, no. 2, pp. 46-56, 2011.

[15] G. Szabo and B. a. Huberman, "Predicting the popularity of online content," Communications of the ACM, vol. 53, no. 8, p. 80, Aug. 2010.

014.