Predicting the Popularity of Online News from Content Metadata

Md. Taufeeq Uddin, Muhammed Jamshed Alam Patwary, Tanveer Ahsan, Mohammed Shamsul Alam
Department of Computer Science and Engineering, International Islamic University Chittagong, Bangladesh
mohamed.taufeeq.uddin@gmail.com, jap_cse@iiuc.ac.bd, tanveer.ahsan@gmail.com,alam_cse@yahoo.com

***Abstract*—Popularity prediction of online news aims to predict the future popularity of news article prior to its publication estimating the number of shares, likes, and comments. Yet, popularity prediction is a challenging task due to various issues including difficulty to measure the quality of content and relevance of content to users; prediction difficulty of complex online interactions and information cascades; inaccessibility of context outside the web; local and geographic conditions; social network properties. This paper focuses on popularity prediction of online news by predicting whether users share an article or not, and how many users share the news adopting before publication approach. This paper proposes the gradient boosting machine for popularity prediction using features that are known before publication of articles. The proposed model shows around 1.8% improvement over previously applied techniques on a benchmark dataset. This model also indicates that features extracted from articles keywords, publication day, and the data channel are highly influential for popularity prediction.**

***Keywords*—Social Media Contents, Popularity Prediction, Before Publication Approach, Machine Learning, Text Mining**

I. INTRODUCTION

The consumption of online news accelerates day by day due to the widespread adoption of smartphones and the rise of social networks. Note that online news content comprises of numerous key properties. For instance, it is easily produced and small in size; its lifespan is short and the cost is low. Such properties make news content more effective to be consumed on social sharing platforms. More interestingly, this type of content can capture the attention of a significant amount of Internet users within a short period of time. As a consequence, researchers focus on the analysis of online news content such as predicting the popularity of news articles, demonstrating the decay of interest over time to understand the world of online news since it has so many practical implications [1].

The prediction of the popularity of online news content has remarkable practical values in many fields. For example, by utilizing the advantages of popularity prediction, news organization [2] can gain a better understanding of different types of online news consumption of users. As a result, the news organization can deliver more relevant and engaging content in a proactive manner as well as the organization can allocate resources more wisely to develop stories over their life cycle. Furthermore, prediction of news content is also beneficial for trend forecasting, understanding the collective human behavior, advertisers to propose more profitable monetization techniques, and readers to filter the huge amount of information quickly and efficiently [1] [3].

978-1-5090-6122-8/16/$31.00

The notion of popularity is often expressed by investigating the number of interactions in the web and social networks, for example, click-through rate, number of shares, likes, and retweets. Tatar et al. [4] demonstrated two types of popularity prediction techniques that are **after publication:** more common technique, which uses features capturing the attention that one content receives after its publication. Higher prediction results are expected in after publication technique since utilization of information about the received attention makes the prediction task easier [1] [5] [6] [7]; **before publication:** relatively challenging and effective technique. This technique uses only content metadata features that are known prior to the publication of contents instead of using features related to the attention that one content receives after contents release. Although the expected prediction accuracy is comparatively low in before publication method as we are using only metadata features rather than original news content [8], the prediction is more desirable as far as it fosters the possibility of decision making to customize the content before the release of content [9]. In this work, we model popularity prediction problem in before publication technique.

Although popularity prediction of web content has tremendous impacts in many areas, popularity prediction task still faces a bunch of major challenges [4] [9]. First, different factors make prediction difficult, for example, the quality of content or relevance of content to users can influence contents popularity. Second, the relationship between events in the real world and content itself are not only difficult to capture but also hard to further feed into the prediction engine. Third, prediction of complex social interactions and information cascades at the microscopic level are extremely challenging. Fourth, the prediction might also be difficult because of the inaccessible content like context outside the web, local and geographical conditions, and situations which influence the population. Last but not least, the prediction may also be hard based on the network properties e.g. the structure of the networks, and the interplay between different layers of the web.

Previously, researchers try to estimate the popularity by predicting whether or not someone shares the news. However, this approach is less informative since we can only identify

users share the news rather than how many users share the news. Hence, this paper proposes an extension to the previous popularity prediction models by predicting the number of shares of news using a novel ensemble learning algorithm, namely gradient boosting machine (**GBM**) [10] in **before publication** setting. In this work, we use a heterogeneous set of metadata features that are known prior to the publication of the article to train the proposed GBM, where the goal is not only to predict whether users share an article or not but also to predict how many users shares the article. Hence, GBM is the ensemble of multiple weak learners or decision trees in which each successive decision trees are built from the prediction residual of the preceding decision trees to form a final highly accurate prediction model. The final prediction model guarantees that it performs much better than the individual performance of each decision tree. Note that the proposed GBM comprises of several good qualities including high popularity prediction accuracy, competitive computational performance both in training and prediction steps, and capability to handle large training dataset.

II. RELATED WORK

Over the last couple of years, researchers conducted several web mining and machine learning studies regarding web content analysis. For instance, Tatar et al. [4] analyzed different types of web content such as online videos, online news and social networking sites. They also reviewed different web content popularity prediction models including both classification and regression models. They further presented *good predictive features such as characteristics of content creators, textual features, and sentiment analysis, and revealed influential factors toward web content popularity*. Gao et al. [3] investigated the arrival process of retweets as well as user activity variation on the retweeting dynamics. They predicted the popularity by modeling the retweeting dynamics using extended reinforced Poisson process model with time mapping process. They, in addition, reduced the effect of user activity variation introducing the Weibo time notation as well as integrating a time mapping process into the proposed model. Castillo et al. [2] provided a qualitative and quantitative analysis of the life cycle of articles' stories demonstrating the interplay between site visitation patterns and social media reactions to articles. Furthermore, they modeled overall traffic of articles by observing social media reactions or attention profile such as *decreasing, steady, increasing,* and *rebounding*.

After publication method is highly popular in popularity prediction research. For instance, Tatar et al. [1] ranked news articles by predicting *user comments* using the linear model on a logarithmic scale and constant scaling model in after publication setting. They outlined that *popularity prediction methods are the good alternative for automatic online news ranking*. Lee et al. [5] inferred the likelihood of the popularity of online content for survival analysis applying Cox proportional hazard regression. They used a set of observable explanatory factors to model and to predict objective metric such as threads lifetime and the number of comments. Petrovic et al. [11] predicted message propagation in twitter using passive-aggressive (PA) learning algorithm in time-sensitive approach. They extracted features related to the author and text of the tweets and various statistics of the tweet itself. Their findings suggested that *automatic retweets prediction performance of PA is as good as prediction performance of humans*. Szabo and Huberman [6] predicted the long-term dynamics of individual submissions e.g. number of views for Youtube videos and the number of votes for Digg stories from early measurements of access of users. They highlighted that Digg stories outdated shortly while Youtube videos were found popular for a long time after their initial submission to the portal.

However, there are only a few studies which followed the challenging *before publication approach* like this study. For example Bandari et al. [9] constructed features from the content of the news articles and its source of publication that were available prior to contents release. They considered four characteristics of the articles: *news source, the category of news, the subjectivity of the language,* and *named entities mentioned in the articles*. They reported that ranges of popularity on social media could possibly be predicted with 84% accuracy using the bagging technique. Arapakis et al. [12] pointed out that *news popularity prediction at cold start is still an open challenge*. They predicted tweet counts and page views using features related to time, news source, genre, Wikipedia, web search and twitter. In their findings, they reported that *the imbalanced class distribution drove the prediction models to bias toward the unpopular articles that concluded the predictions not useful in realistic scenarios*. Fernandes et al. [8] proposed proactive intelligent decision support system for online news articles that predicted whether user shares articles or not analyzing features known before publication such as keywords, digital media content, and earlier popularity of news referenced in articles. They achieved 73% popularity prediction accuracy on Mashable news dataset via random forests algorithm adopting rolling windows evaluation strategy.

III. GBM FOR NEWS POPULARITY PREDICTION

The goal of this work is to predict whether a news article may share or not by users as well as the total count of shares in the realistic setting using GBM algorithm in both classification and regression settings. In training step of this approach, heterogeneous set of metadata features extracted from articles is fed to GBM to build the desired prediction model. The trained GBM model is then used in prediction step to carry out popularity prediction task. The rest of this section covers a brief description of the key idea, and regularization techniques of GBM for classification and regression.

A. Key Idea

GBM [13] is an ensemble learning algorithm that is the combination of gradient-based optimization and boosting. GBM produces strong prediction model by combining multiple weak prediction models in which weak models are created by sequentially applying to the incrementally changed dataset.

Optimization based on the gradient in GBM utilize the gradient computations to minimize the cost function of a model with respect to training dataset while boosting additively gathers an ensemble of weak models to build the prediction model for popularity prediction challenge. In short, the main idea beneath GBM is to build a series of simple and probably inaccurate decision trees or weak models successively from the prediction residuals of the preceding decision trees and combine them to construct a final highly accurate prediction model.

B. Regularization

Generalization capability of GBM is one of the crucial concern [14]. A number of parameters can contribute towards reducing the effects of overfitting by controlling learning rate and/or by introducing randomness into GBM. For instance, the smaller values of learning rate v such as $v \leq 0.1$ can generally ensure better generalization and performance on the test dataset. Overfitting can also be eliminated by fitting weak models on a subsample or constant fraction e.g. $0.5 \leq fraction \leq 0.8$ of the training dataset at random with no replacement. Furthermore, to reflect generalization, we can use a large number of trees or the boosting iteration M, and control J, by picking the value of J in between 4 and 8.

IV. MASHABLE NEWS DATASET

Online news popularity prediction (**Mashable news**) dataset [8] is publicly available at `http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity`, which aims to predict the future popularity of news articles using information that are known before the release of news articles. Mashable news dataset consists of 58 heterogeneous features about the associated statistics of the original news articles released by Mashable (`www.mashable.com`) during a two years period from January $7, 2013$ to January $7, 2015$. Fernandes et al. crawled news articles from Mashable website, and then they discarded special occasion and very recent articles e.g. < 3 weeks from the crawled articles.

More precisely, they extracted 47 features from HTML code and classified them into 4 different categories such as *number, ratio, bool*, and *nominal*. The unbounded numeric features like the number of words in the article were scaled by logarithmic transformation as well as they transformed the nominal features with the common *1-of-C* encoding. They also extracted the statistical summary of the number of shares of all Mashable links cited in articles that were known before the release of articles. Three types of keywords such as worst, average and best were captured by ranking all articles keyword average shares that were also known before release.

Additionally, they extracted a bunch of natural language processing features such as closeness to top latent Dirichlet allocation (LDA) [15] topics, title subjectivity, the rate of positive and negative words and title sentiment polarity by using LDA to compute relevant topics as well as to measure the closeness of current article to the previously computed topics. Sentiment polarity and subjectivity scores were also computed by applying the pattern web mining module [16].

V. EXPERIMENTS AND RESULTS

The proposed GBM prediction model was evaluated on Mashable news dataset using $5 - fold$ cross-validation strategy. To benchmark the proposed GBM against previously applied algorithm, namely random forests, we also reproduced random forests (**RF**) from reference [8] adopting 5-fold cross validation. Each of the experiments was run for 20 times with different random seeds, and then we averaged over 20 different experimental runs to achieve the final results. For predicting whether users share a news article or not, we modeled the popularity prediction problem as binary classification problem that is $Popular$ vs. $Unpopular$. In the case of binary classification, receiver operating characteristics (ROC) curve was produced to demonstrate the performance of the models. Note that ROC is created by plotting sensitivity against one minus specificity i.e. $1 - specificity$ at numerous threshold values. The larger value of the performance metric area under the ROC curve (**AUC**) indicates the higher popularity prediction accuracy.

As mentioned earlier, the experimental set up was binary classification problem in which the goal was to classify whether an article was popular or unpopular by predicting whether users share that article or not. We defined the popularity of an article based on a decision threshold D. For instance, when an article is shared more than 1400 times e.g. $D \geq 1400$, we labeled the article as $Popular$; otherwise, we labeled the article as $Unpopular$ as suggested in [8]. We considered AUC as the primary evaluation metric as far as AUC is the most suitable metric as AUC is independent of the threshold value as well as AUC calculates discrimination power of the models very efficiently.

On the other hand, we modeled the popularity prediction problem as regression problem during the estimation of popularity via predicting the number of shares of the news article by users. Note that we took the logarithm of the original number of shares of news to use as the prediction label to train and test using GBM and RF regressors. During performance evaluation of the regression models, we evaluated the models by measuring the mean absolute percentage error between the predicted shares and logarithm of the original number of shares (**MAPE_LOG**), and between the exponents of predicted shares and the original number of shares (**MAPE**).

For training and validating the proposed GBM, we set *logistic regression*, and *linear regression* as objective functions during binary classification and regression, respectively, the value of the learning rate $v = 0.001$ to reflect better generalization, the size of the tree $J = 8$, and subsample size $fraction = 0.8$. Finally, the number of boosting iteration M was selected using 5-fold validation approach. Figure 2 shows the performance of RF and GBM classifiers. Notice that the best AUC value obtained on Mashable news dataset was **74.5%** using GBM. It can be seen that GBM performed **1.8%** better than widely used machine learning algorithm RF which generated previous benchmark results on Mashable news dataset. Although the obtained result was around **74-75%** discrimination level which was far away from being

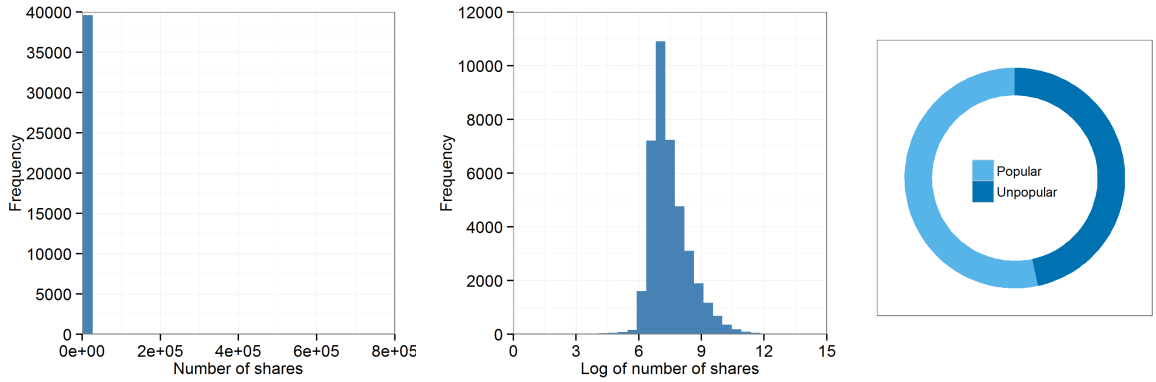

Fig. 1. From left to right, distribution of the popularity prediction label e.g. original number of shares, the logarithms of the number of shares, and proportion of the popular and unpopular article classes of Mashable news dataset

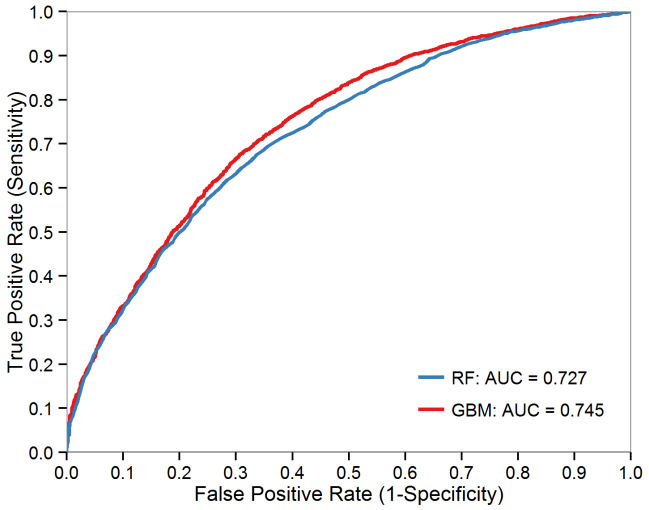

Fig. 2. Comparison of the AUC produced from RF and GBM

TABLE I
PREDICTION BENCHMARK OF NUMBER OF SHARES OF NEWS ARTICLES. THIS TABLE DISPLAYS MAPE_LOG, MAPE, AND THEIR STANDARD DEVIATION (**SD**) RESULTED FROM RF AND GBM APPLYING ON MASHABLE NEWS DATASET

Model	MAPE_LOG	MAPE_LOG SD	MAPE	MAPE SD
RF	8.39 %	0.0083	73.41 %	0.0986
GBM	**8.11 %**	0.0073	**69.42 %**	0.0989

perfect, performance was still interesting since we only fed metadata features for prediction task following before publication technique. As we obtain this prediction results using some meta features or statistical features before publication of the news, we can, therefore, feed this results for further modification of the articles.

Table I shows the performance of RF and GBM during the prediction of the count of shares of articles. From Table I, it can be indicated that GBM was better than RF since MAPE_LOG and MAPE generated from GBM were 8.11% and 69.42%, respectively, that were relatively smaller than the MAPE_LOG and MAPE generated from RF.

GBM uses *gain* [17] to estimate the contribution of each feature to the prediction model. GBM takes each gain of each feature of each tree and computes mean per feature to provide a vision of the entire prediction model. We measured the relative importance scores of features using GBM for both predicting popularity and number of shares of news for which we trained GBM using 35679 news articles. Figure 3 highlights the relative importance scores or $Gain * 100$ of the top 20 features; x-axis represents the relative feature importance scores and y-axis represents the 20 features from Mashable news dataset. In can be observed from Figure 3 that features related to the summary statistics of contents keywords, the summary statistics of the number of the shares of the referenced articles, tokens, links, data channel types whether it is entertainment or technology channels, and publication day had strong significance for predicting popularity and number of shares of online news. Moreover, natural language processing features such as closeness of target article to different LDA topics e.g. 0, 1, 2, rate of positive and negative words, text subjectivity, and global sentiment polarity encapsulated high discrimination power towards popularity prediction.

VI. CONCLUSION AND FUTURE WORK

This paper introduces and implements GBM to tackle the challenge of classifying popular news articles from unpopular articles by measuring the count of shares in before publication approach. Our findings suggest that GBM is able to predict popularity with a decent prediction rate using only statistical features associated with original news articles without using the original content of news articles or after publication attention. GBM also outlined discriminative and useful metadata features such as the statistical summary of keywords, the earlier popularity of articles referenced in articles, natural language processing features, and publication time. Future work will include, first, the exploration of more advanced features regarding content like trend analysis. Second, the evaluation of the prediction model on more complex and

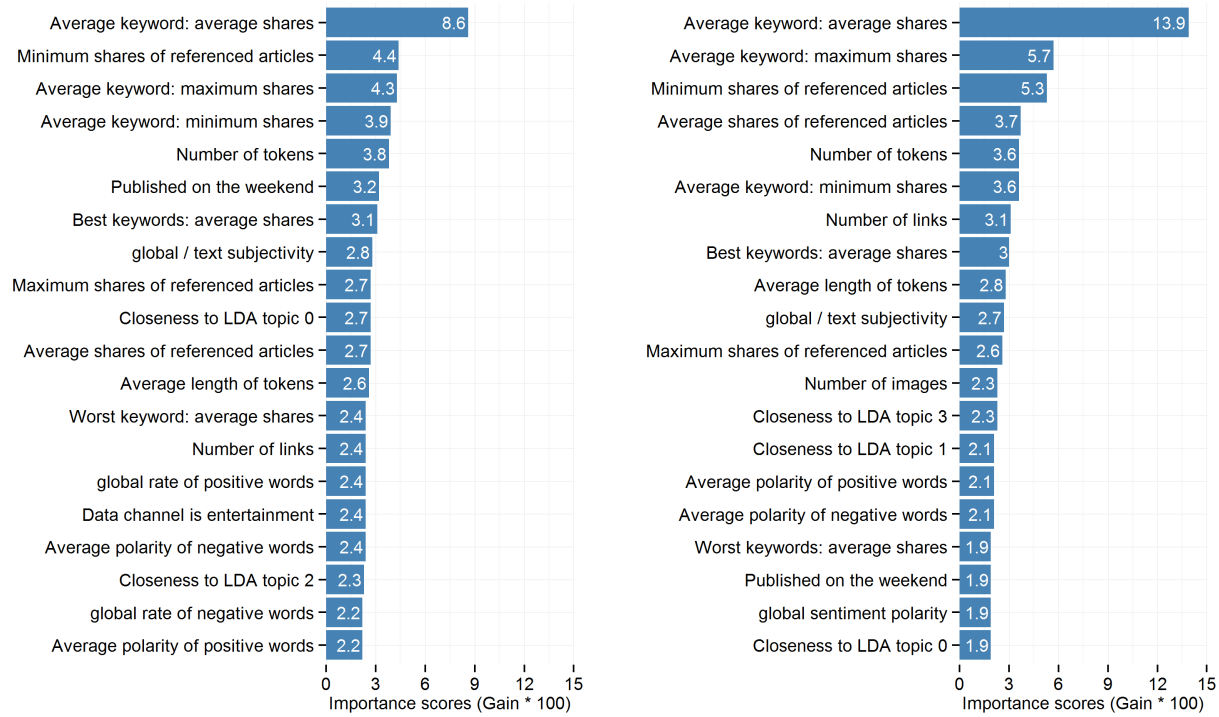

Fig. 3. Top 20 features based on the importance scores of features, measured via GBM, for predicting the, from left to right, popularity of online news, and number of shares of online news

more unbalanced popularity prediction datasets. Third, the comparison of the model with many other state-of-the-art techniques.

REFERENCES

[1] A. Tatar, P. Antoniadis, M. Amorim, and S. Fdida, "From popularity prediction to ranking online news," *Social Network Analysis and Mining*, vol. 4, no. 1, 2014. [Online]. Available: http://dx.doi.org/10.1007/s13278-014-0174-8

[2] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck, "Characterizing the life cycle of online news stories using social media reactions," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ser. CSCW '14. New York, NY, USA: ACM, 2014, pp. 211–223. [Online]. Available: http://doi.acm.org/10.1145/2531602.2531623

[3] S. Gao, J. Ma, and Z. Chen, "Modeling and predicting retweeting dynamics on microblogging platforms," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ser. WSDM '15. New York, NY, USA: ACM, 2015, pp. 107–116. [Online]. Available: http://doi.acm.org/10.1145/2684822.2685303

[4] A. Tatar, M. de Amorim, S. Fdida, and P. Antoniadis, "A survey on predicting the popularity of web content," *Journal of Internet Services and Applications*, vol. 5, no. 1, 2014. [Online]. Available: http://dx.doi.org/10.1186/s13174-014-0008-y

[5] J. G. Lee, S. Moon, and K. Salamatian, "Modeling and predicting the popularity of online contents with cox proportional hazard regression model," *Neurocomputing*, vol. 76, no. 1, pp. 134–145, 2012.

[6] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[7] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini, "A peek into the future: Predicting the evolution of popularity in user generated content," in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, ser. WSDM '13. New York, NY, USA: ACM, 2013, pp. 607–616.

[8] K. Fernandes, P. Vinagre, and P. Cortez, *Progress in Artificial Intelligence: 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11, 2015. Proceedings*. Cham: Springer International Publishing, 2015, ch. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News, pp. 535–546.

[9] R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity," *CoRR*, vol. abs/1202.0332, 2012. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1202.html#abs-1202-0332

[10] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, 2000.

[11] S. Petrovic, M. Osborne, and V. Lavrenko, "Rt to win! predicting message propagation in twitter." in *ICWSM*, L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds. The AAAI Press, 2011. [Online]. Available: http://dblp.uni-trier.de/db/conf/icwsm/icwsm2011.html#PetrovicOL11

[12] I. Arapakis, B. Cambazoglu, and M. Lalmas, "On the feasibility of predicting news popularity at cold start," in *Social Informatics*, ser. Lecture Notes in Computer Science, L. Aiello and D. McFarland, Eds. Springer International Publishing, 2014, vol. 8851, pp. 290–299. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-13734-6_21

[13] M. M. Cliff Click, Jessica Lanford and V. Parmar, "Gradient boosted models with h2o's r package," February 2015: Second Edition, published by H2O.ai, Inc. 2307 Leghorn Street Mountain View, CA 94043, Inc. Available: https://leanpub.com/gbm/read. [Online]. Available: https://leanpub.com/gbm/read

[14] "Gradient boosting," wikipedia, Wikimedia Foundation, Inc. , Web. Last Accessed: 26 August 2015. [Online]. Available: https://en.wikipedia.org/wiki/gradient_boosting

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[16] T. De Smedt, L. Nijs, and W. Daelemans, "Creative web services with pattern," in *Proceedings of the Fifth International Conference on Computational Creativity*. Citeseer, 2014.

[17] tianqi chen, kailong chen, and tong he, "Xgboost," 2015, http://mloss.org/software/view/543/.