doi: 10.36689/uhk/hed/2021-01-025

Suitability of Machine Learning Methods for Prediction of Popularity on Social Media in Comparison of Different Data Sets

Jan HRUŠKA

University of Hradec Králové, Hradec Králové, Czech Republic, hruskja3@uhk.cz

Abstract: The growing popularity of various social media and the use of neural networks in recent years has brought predicting opportunities in various sectors. Social networks, in conjunction with neural networks, are often used in healthcare (prediction of whether a disease occurs or symptoms return) or business (what will be the profit or error rate of the products) and are playing an important economic and marketing change in the 21st century. The aim of the presented project is to contribute to a deeper understanding of which machine learning method to use for which data set in order to ensure the best possible prediction success across different industries. A total of 10 methods were used and the success of a total of 13 data files was tested with the help of specialized software for machine learning and the Python programming language. Of all the methods tested, the highest success rate on most datasets was with the Random Forrest method. The success rate ranged from 58.38% to 98.65%. Out of the total number of 13 datasets, the Random Forrest method was 5 times the best in accuracy.

Keywords: machine learning; prediction; efficiency

JEL Classification: C53

1. Introduction

The constant development of social media and interactive behavior among users often causes user-generated content to spread rapidly over the Internet. The popularity of social media and its content allows us to understand these activities, which have an impact on social, economic, and government activities. Modeling and predicting the popularity of online content is an important area of social media research and can have a positive impact in many ways on public administrations, organizations, businesses, and online security. For example, it can support crisis management by knowing the effects of natural disasters, terrorism, or crime (Chen et al., 2019). In business, this prediction can help analyze current trends, and concerns among users and provide valuable information for risk modeling and highlight potential benefits (Cerchiello et al., 2017).

However, the prediction is not a trivial task. The amount of content generated on the Internet and by users on social media is huge. Methods are also constantly evolving and often have many possible settings. For example, there are hundreds of millions of active users per month on Facebook and Twitter alone, and linking their relationships is a very complicated issue (Ballings et al., 2016; Moro et al., 2016). The data is also often not structured and there are

many informal expressions. Last but not least, there are many factors that affect the success of a prediction, such as the size of a data file, the number of variables, the correlation between variables, and so on. In the case of social media, also text content, user type and time of addition, length of the post, selected image or video, positive or negative attunement, appropriateness and length of the title, number of responses during the first few hours, and many other variables (Szabo & Huberman, 2008; Vazquez et al., 2014).

Popularity prediction studies primarily use property-based methods (Tsur & Rappoport, 2012; Weng et al., n.d.). These methods and models are highly dependent on choosing the right functions, which can be time-consuming and requires expertise. If inappropriate properties are selected, the model will be inaccurate and fragile. Initial studies focused on simple regression models, but some models in recent years using the popularity of stochastic modeling (Mishra et al., 2016). Quite often, however, these models use only time series prediction and ignore other valuable information to predict popularity.

The popularity of social media is increasing attention to online content in terms of the number of shares, tweets, views, or "likes." Popularity prediction is focused on predicting these variables using various properties, such as a text post, event, video, or image. Existing approaches in these prediction models fall into four categories: pure property prediction, time series analysis, cascade process analysis, and in-depth machine learning (Chen et al., 2019).

The prediction uses properties to extract variables such as surface variables (such as publication time, content length, title length, or number of images and links in a post), cumulative variables (number of articles published simultaneously), text properties (for example in the text), semantic properties (possibly selected locations, people or organizations and their current popularity) or physical characteristics (temperature, weather) (Bandari et al., 2012; KR et al., 2018).

There are several possible neural networks for machine learning or deep neural networks that can be used to predict popularity. But the basic neural network has become a repetitive neural network (RNN). Extension models for this basic neural network are LRCN (Long-term Recurrent Convolutional Networks) (Donahue et al., 2014), LSTM (Serban et al., 2015), and GRU (Gated Recurrent Unit) (Chung et al., 2014). However, neural networks assume changes monotonic for the entire sequence, and these factors can be more complex on social media, in addition to variables such as sharing time (business hours, evenings, weekends), and can be difficult to capture using classical repetitive neural networks (RNNs).

However, fewer studies pay attention to research focused on the implementation of predictive systems, which can be effectively used to predict the development of the paper before its publication. A system capable of predicting the impact of a published individual impact of a paper can provide a valuable advantage in deciding on communication through social media, in customizing the promotion of products and services. Based on the outcome of these predictive systems, advertising managers could make decisions. Data mining provides an interesting approach to extracting predictive knowledge from raw data. However, most studies focus on reactively evaluating what users say through a series of responses, posts, or tweets.

1.1 Benefits of Social Media Marketing

- Large audience – Facebook has more than 2.7 billion monthly active users. Twitter has more than 320 million monthly active users, Instagram more than 500 million, and this makes social media a great opportunity for creators around the world and across all industries.
- Free Creation - The beginning of creating any marketing strategy is free on all major social media. Of course, paid marketing tools guarantee that the number of potential customers will increase drastically, but marketing can be done completely free of charge.
- Encourages sharing - Perhaps the most unique benefit of social media is the ability to get help from their followers. People share things with their networks, from photos and recipes to interesting articles and hot deals. Unlike other forms of Internet marketing, such as sites and paid advertising, content on social media is often shared. Followers can share with their followers, who then share with their followers, giving a wider reach (at a lower cost) than a traditional marketing campaign (Ding et al., 2019).
- Stronger brand loyalty - In addition to increasing brand reach, social media can increase brand loyalty. A study by The Social Habit shows that 53% of Americans who follow certain brands on social networks are more likely to be loyal to those brands. Social media is more than a selling point (Svobodová & Hedvičáková, 2018). Instead, it facilitates two-way communication that allows you to build meaningful relationships with existing and potential customers. This makes them more confident in their decision to trust their business and encourages them to choose your brand in the future.

1.2. Disadvantages of Social Media Marketing

- Negative Feedback - The biggest disadvantage of social media marketing is that negative feedback can be devastating. Social media users can write whatever they want. A happy customer can leave perfect reviews, but unhappy customers can leave angry, hateful comments, and everyone on social media sees it.
- Embarrassment - With various marketing posts and campaigns, it's easy to make a mistake or just use a bad ambiguous word that can escalate with a large audience.
- Time consuming - Creating and maintaining an interactive marketing profile on most social media can take a long time and effort. If the social media team is small with limited resources, it can be difficult to maintain good results.

In recent years, a completely new marketing channel has opened up for marketing specialists. Companies now have a unique opportunity to respond very quickly and to some extent personally to customer opinions, questions or complaints about defective goods. Advertising through word of mouth (WoM) is extremely attractive for them. It is a free form of advertising, where customers tell others that they like a certain company, product or service. People have recommended products and services since ancient times and did not need the Internet or computers. In recent years, however, we have increasingly encountered electronic verbal transmission. If information is spread extremely quickly through personal transmission, one can even speak of viral marketing. This term was allegedly first used in

1996 in the business magazine Fast Company and got its name due to its similarity to the spread of various diseases. More specifically, it can be defined as "electronic word of mouth, in which a given form of marketing message concerning a company, brand or product is transmitted at an exponentially increasing rate, often via social media." (Ding et al., 2019).

2. Methodology

The aim of this work is to use analysis and testing to reveal which method is most suitable for prediction on the data obtained by the programmed algorithm (shown below) and also in general and which methods are suitable for different datasets according to the number of variables in the dataset and its size. The data provide information about articles on the Internet and with their help, the author tries to predict the success of the article in terms of the number of shares (available information listed directly with the article). The Python programming language and the Selenium web browsing library were used to program the algorithm. IBM SPSS Statistics software was used for the first data analysis and WEKA (Waikato Environment for Knowledge Analysis), Keras software was used for the creation and testing of neural networks. In the case of Keras software, this is another Python library. The expected number of saved articles was just over a thousand before starting data acquisition, but this number was increased to almost 17 thousand due to the acceleration of the algorithm. The analyzed articles are from the website mashable.com, which unfortunately continued to cancel the display of shares during this work, there are other sites, which list the number of shares, but are often focused on one specific area such as healthcare (psychcentral.com) or social media marketing (socialmediaexaminer.com). Another suitable site for this research would be thenextweb.com, but for simplicity, article variables were loaded only from the Mashable page. PyCharm software was used for programming. Random Forest, Naive Bayes, and K Nearest Neighbors methods will be programmed to analyze the success of the prediction. The success will then be compared in tables with other methods from WEKA software.

2.1. Datasets

All datasets are freely available datasets from the site (https://github.com/renatopp/arff-datasets/tree/master/classification). The only data set obtained by the author using the programmed algorithm is listed in Table 1 as "Articles". The variables obtained include:

- TitleWords = number of words in the title of the article
- TitleCharacters = number of characters in the article title
- TotalWordsInArticle = number of words in the whole article
- TotalCharactersInArticle = number of characters in the article title
- TotalParagraphs = total number of paragraphs
- AverageWordLengthInTitle = average length of the word in the article title
- AverageWordLengthInArticle = average word length in the whole article
- NumberOfImages = number of images in the article
- NumberOfHrefs = number of links in the article

- PositiveWordsInArticle = number of positive words in the article
- NegativeWordsInArticle = number of negative words in the article
- PositiveDividedByNegativePolarityInArticle = number of positive words divided by the number of negative words in the article (article polarity)
- PositiveDividedByAllWordsInArticle = number of positive words divided by all words in the article
- NegativeDividedByAllWordsInArticle = number of negative words divided by all words in the article
- PositiveWordsInTitle = number of positive words in the article title
- NegativeWordsInTitle = number of negative words in the article title
- PositiveWordsInTitleDividedWordsInTitle = number of positive words in the title divided by the number of words in the article title
- NegativeWordsInTitleDividedWordsInTitle = number of negative words in the title divided by the number of words in the article title
- WordsInFirstParagraph = number of words in the first paragraph
- CharactersInFirstParagraph = number of characters in the first paragraph
- PositiveInFirstParagraph = number of positive words in the first paragraph
- NegativeInFirstParagraph = number of negative words in the first paragraph
- PositiveInFirstDividedByWordsInFirst = number of positive words in the first paragraph divided by the number of words in the first paragraph
- NegativeInFirstDividedByWordsInFirst = number of negative words in the first paragraph divided by the number of words in the first paragraph
- DayPosted = day of the month the article is added (1-31)
- MonthPosted = month when the article is added
- YearPosted = year when the article is added

The last variable examined:

- NumberOfShares = total number of shares on all social networks combined

To obtain the number of positive and negative words, the entire analyzed text was compared word for word with the external documents positiveWords.txt and negativeWords.txt. These documents contain positive and negative English words. These lists are freely available at https://positivewordsresearch.com/sentiment-analysis-resources/. After downloading the files, it was necessary to go through the documents manually and remove the words for which the polarity is questionable. For example, the word "joke" was listed as negative and so on.

In the case of other datasets, there is a wider range of predictions. There are datasets for predicting whether the symptoms of a disease will return (dataset "Diabetes" and "Breast Cancer"), datasets for predicting quality (Glass, Wine Quality, Soybean, Ionosphere), as well

as predicting the type of plant by leaves ("Iris"), or whether the person is a Republican or a Democrat ("Votes") or whether the customer is credible ("German Credit").

3. Results

For the programmed methods, the prediction success for the "Articles" dataset was 51.33% for the Naive Bayes method, 58.38% for the Random Forest method, and 55.64% for the K Nearest Neighbor method. Then, the following table 1 was expanded with other data sets listed in the methodology and machine learning methods available in the WEKA software.

Table 1. Comparison of methods for selected datasets and their prediction success.

Method\Dataset	Iris	Weather	Contact Lenses	Diabetes	Breast Cancer	Glass	Wine Quality	Labor	Votes	German Credit	Articles	Soybean	Ionosphere
Number of variables	4	4	5	8	9	9	11	16	16	20	27	35	35
Number of possible prediction outputs	3	2	3	2	2	7	10	2	2	2	2	19	2
Number of instances	150	14	24	768	286	214	3429	57	435	700	16784	683	351
Naive Bayes	94.12%	60.00%	37.50%	78.54%	71.13%	49.32%	45.28%	**94.74%**	91.22%	76.37%	51.33%	90.52%	82.35%
Random Forest	96.08%	**80.00%**	50.00%	78.54%	71.10%	**78.08%**	**63.46%**	89.47%	**98.65%**	**76.89%**	**58.38%**	93.10%	80.67%
K Nearest Neighbor	96.08%	40.00%	50.00%	72.80%	72.16%	67.12%	55.57%	78.95%	91.90%	73.11%	55.64%	88.79%	**91.60%**
Logistic regression	98.02%	60.00%	**62.50%**	79.31%	68.04%	54.79%	53.00%	89.47%	97.97%	75.63%	53.49%	**93.53%**	85.71%
RepTree	92.16%	60.00%	37.50%	75.48%	65.98%	57.53%	50.26%	84.21%	97.97%	72.69%	53.87%	80.17%	84.87%
Support Vector Machines	**100.00%**	60.00%	37.50%	68.20%	65.98%	64.38%	53.69%	89.47%	96.62%	71.85%	54.92%	84.91%	89.08%
MultilayerPerceptron	98.04%	60.00%	50.00%	80.08%	**74.23%**	57.53%	54.63%	89.47%	97.30%	**76.89%**	53.10%	55.60%	81.51%
RBFClassifier	94.12%	60.00%	37.50%	77.01%	69.07%	50.68%	52.40%	78.95%	96.62%	75.63%	54.07%	44.83%	89.92%
Stochastic Gradient Descent	NA	60.00%	NA	**80.84%**	71.13%	NA	NA	84.21%	97.97%	**76.89%**	53.66%	NA	83.19%
Fuzzy Unordered Rule Induction Algorithm	96.08%	40.00%	37.50%	80.46%	65.98%	64.38%	52.92%	89.47%	97.40%	73.11%	56.05%	93.10%	88.24%

The table shows selected methods and datasets along with the prediction success. The grayed out methods are those methods that were programmed in the Python programming language where success was tested. Other methods were run in WEKA software. The highlighted data set "Articles" is described above in the article, which was obtained using the created algorithm. The method with the highest accuracy for the given dataset is highlighted in bold. Among the selected datasets, Random Forest proved to be the most universal method. This method also has the best success for predicting popularity on social networks in this work. This success rate is 58.38%, which means that if an article structure is inserted into a method, this method can indicate with the stated accuracy whether the article will be shared more than the average.

4. Discussion

The worldwide spread of social media has led to exponential growth in Internet users, leading to a whole new environment for customers, the exchange of ideas and feedback on products and services. (Bandari et al., 2012). Thanks to rapid development, social media may become the most important media channel for brands and clients in the near future (Szabo & Huberman, 2008). Companies soon realized the potential of using online social services to influence customers and include social media marketing and other various surveys. Nowadays, a number of companies, social networks and individuals already use various predictive models, and this work can point to the fact of which methods are suitable for which issues and for which platform.

Social media websites such as Twitter, Facebook, and YouTube are based on human interaction and user-generated content. This leads to the creation and exchange of huge amounts of user-generated content in one-to-many communications. Social media-oriented people tend to publish texts, audio, or video related to their lives, thereby demonstrating their tastes and opinions (Ding et al., 2019).

The data available on social media platforms provide a wealth of information on human behavior and social interaction. Social media provides data to understand needs and opinions. By analyzing what is available on social media, it is possible to identify important personality traits, i.e. traits or characteristics specific to humans, which can then be used not only to describe their personality. (Lima & Castro, 2014). All this information can be used as a further extension of various datasets and after appropriate modification of the data can be inserted into the methods presented in this work and further refine predictions in many different areas (for example, profit generation, detection and prediction of user mood, disease spread prediction, prediction election result and much more).

The limits of this work include focusing on a smaller number of websites providing articles and the number of shares of their article. The limit of this work may also include focusing only on the years 2013 and 2014.

5. Conclusions

Of the methods tested in this project, including Naive Bayes, Random Forest, K Nearest Neighbor and subsequently also logistic regression, decision tree, support vector machines,

and other modifications, the success rate ranged from 37.5% to 100%, where the Random Forest method was the most successful (in 5 of 13 datasets). The worst methods in overall accuracy were methods RepTree, RBFClassifier, and Fuzzy Unordered Rule Induction Algorithm.

Today, businesses and organizations are increasingly inclined to take advantage of the ubiquitous impact of online social media such as Facebook, Twitter, and Instagram. These companies and their campaigns reach different categories of users very quickly, as many people spend most of their time on online social media (Ducange et al., 2019).

However, effective prediction of the number of shares on social media will be related to a large number of other variables such as user characteristics (age, nationality, frequency of sharing, number of social networks used, ...) or domain characteristics (who wrote the article, what is the site traffic, whether they have paid advertising, how much they spend on advertising and the like).

Overall, the predicting is a popular but complex topic. This fact is also pointed out by the fact that for almost every data set there was the most accurate method and it is, therefore, necessary to think carefully about the prediction method used depending on how large a data set is available and how many prediction outputs are possible.

Acknowledgments: The work was supported by the internal project "SPEV – Economic Impacts under the Industry 4.0 / Society 5.0 Concept ", 2021, University of Hradec Králové, Faculty of Informatics and Management, Czech Republic. I would also like to thank doctoral student Martin Matějíček for his help in sorting and preparing data.

References

Ballings, M., Van den Poel, D., & Bogaert, M. (2016). Social media optimization: Identifying an optimal strategy for increasing network size on Facebook. *Omega*, *59*, 15–25. https://doi.org/10.1016/j.omega.2015.04.017

Bandari, R., Asur, S., & Huberman, B. A. (2012). The Pulse of News in Social Media: Forecasting Popularity. ArXiv:1202.0332 [Physics]. http://arxiv.org/abs/1202.0332

Cerchiello, P., Giudici, P., & Nicola, G. (2017). Twitter data models for bank risk contagion. *Neurocomputing*, *264*, 50–56. https://doi.org/10.1016/j.neucom.2016.10.101

Chen, G., Kong, Q., Xu, N., & Mao, W. (2019). NPP: A neural popularity prediction model for social media content. *Neurocomputing*, *333*, 221–230. https://doi.org/10.1016/j.neucom.2018.12.039

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. ArXiv:1412.3555 [Cs]. http://arxiv.org/abs/1412.3555

Ding, K., Wang, R., & Wang, S. (2019). Social Media Popularity Prediction: A Multiple Feature Fusion Approach with Deep Neural Networks. In *Proceedings of the 27th ACM International Conference on Multimedia* (pp. 2682–2686). https://doi.org/10.1145/3343031.3356062

Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., & Darrell, T. (2014). Long-term Recurrent Convolutional Networks for Visual Recognition and Description. ArXiv:1411.4389 [Cs]. http://arxiv.org/abs/1411.4389

Ducange, P., Fazzolari, M., Petrocchi, M., & Vecchio, M. (2019). An effective Decision Support System for social media listening based on cross-source sentiment analysis models. E*ngineering Applications of Artificial Intelligence*, *78*, 71–85. https://doi.org/10.1016/j.engappai.2018.10.014

Lima, A. C. E. S., & de Castro, L. N. (2014). A multi-label, semi-supervised classification approach applied to personality prediction in social media. *Neural Networks*, *58*, 122–130. https://doi.org/10.1016/j.neunet.2014.05.020

Mishra, S., Rizoiu, M.-A., & Xie, L. (2016). Feature Driven and Point Process Approaches for Popularity Prediction. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16* (pp. 1069–1078). https://doi.org/10.1145/2983323.2983812

Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: *A data mining approach. Journal of Business Research*, *69*(9), 3341–3351. https://doi.org/10.1016/j.jbusres.2016.02.010

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2015). Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. ArXiv:1507.04808 [Cs]. http://arxiv.org/abs/1507.04808

Svobodová, L., & Hedvičáková, M. (2018). Factors Determining Optimal Social Media Network Portfolio for Accounting Firms: The Case of the Czech Republic. In S. A. Al-Sharhan, A. C. Simintiras, Y. K. Dwivedi, M. Janssen, M. Mäntymäki, L. Tahat, I. Moughrabi, T. M. Ali, & N. P. Rana (Eds.), *Challenges and Opportunities in the Digital Era* (Vol. 11195, pp. 425–435). Springer International Publishing. https://doi.org/10.1007/978-3-030-02131-3_38

Szabo, G., & Huberman, B. A. (2008). Predicting the popularity of online content. ArXiv:0811.0405 [Physics]. http://arxiv.org/abs/0811.0405

Tsur, O., & Rappoport, A. (2012). What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining - WSDM '12*, 643. https://doi.org/10.1145/2124295.2124320

Vázquez, S., Muñoz-García, Ó., Campanella, I., Poch, M., Fisas, B., Bel, N., & Andreu, G. (2014). A classification of user-generated content into consumer decision journey stages. *Neural Networks: The Official Journal of the International Neural Network Society*, *58*, 68–81. https://doi.org/10.1016/j.neunet.2014.05.026

Weng, L., Menczer, F., & Ahn, Y.-Y. (2014). Predicting Successful Memes Using Network and Community Structure. In *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1). Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14530