Predicting rating of business on Yelp platform using Ensemble training and Graph Convolutional Network

Quang Pham
Undergraduate Student
Temple University, Department of Science and Technology
tug47906@temple.edu

Abstract

Social media platforms have turned into a rating platform for many businesses these recent years. Social media users, who also are customers received services from a business, can affect to the popularity or reputation of the business through sharing information through the social networks (P Aula,2010). With the ability to connect millions of users, Social media give users a chance to influence many businesses success level and also change the behaviors of both customers and businesses (E Qualman,2010). Because of the reason above, businesses should pay effort to understand customers' behavior in order to find potential group of customers to focus on (JP Peter,1999). In this report, I will demonstrate the relationship between the rating of restaurants on Yelp and the profile of users who have rated it. I will also show methods to aggregate information from both users and business on Yelp to reduce the complex of social media data (J Law,2004) and to build a good predicting model for the rating of business on Yelp. For the predicting part, the two models I choose are Ensemble classifier model using voting rules and Graph Convolutional Neural Network.

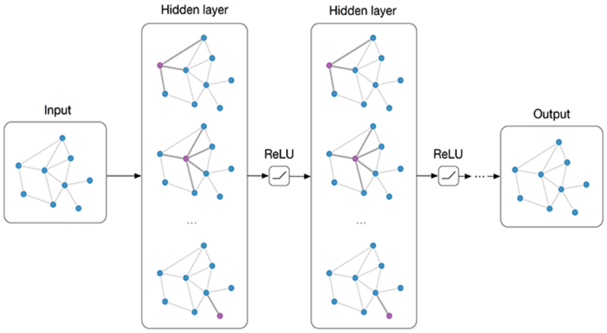
1 Introduction

Social media has a significant impact in our society and the power to connects millions of users with different backgrounds around the world (P Levinson,2009). Going with the huge development of technologies and portable devices, social media gives users many convenient ways to share thought and emotion on different fields of our society. One important element of user activity that valuable is the feedback on services, business and product (JA Martillia, 1997). Because of that reason, many platforms have been developed to provide a directly interaction between users and businesses, such as Yelp. Yelp is a social media platform gather people interested in food related businesses, which mostly used on mobile device. On Yelp, will can find many information of each business such as location, menu, price range and most important the reviews and rating of restaurant. This information was conducted by users on Yelp, also customer of business in the real life. When a customer tried the service of restaurant, he or she can write reviews and give the restaurant a rating score, which scaled from 1 to 5. The higher star a business on Yelp have, the higher quality and reputation the restaurant could have. Since the appearance of Yelp, the success of many businesses in food industry are affected by the platform (M Luca, 2016). Looking into the root of the problem, we are trying to understand how the profiles of the users could affect the rating of business on Yelp. Understanding this relationship could help businesses improve the qualities and benefits by targeting the potential group of customers and providing high-demand services (J Huang, 2014). One problem to notice on Yelp platform as well as other social medias is the credibility of fake information, since fake reviews can affect the reputation of business on Yelp a lot (M Luca,2016). In order to approach this problem, I performed analysis on datasets of business and users on Yelp and tried to find some patterns of relationship among features of datasets. I will also use the Graph convolutional neural networks- a algorithm allow us to reduce the complexity of big network but still remain high amount of data by gather information of each node's neighbors and combined it into an input for the neural networks (M Defferrard, 2016). The main object of my model is to predict the rating of restaurants on Yelp, which could be in range from 1 to 5. To compare with the results of using both information of restaurants and users to predicting the rating, I will use the Ensemble training model to predict rating with the limitation of businesses data.

2 Related Work

Three are many related papers on the used social media data to helped solving challenging problems in different fields, especially when we have many challenges and opportunities from social media (AM Kaplan,2010). Twitter data has been used to detect social patterns of mental illness patients and to find symptoms of mental illness can be detected through tweet activity log such as sleep deprivation and the usage of negative words (G Coppersmith, 2014). Social media data is used to find political opportunities and social movement such as feminism or nationalism (C Shirky,2011). In businesses research, there are proofs so that larges US companies can use Twitter and other social media to gain business value by targeting right group of customers (MJ Culnan, 2010). Social media marketing also has a huge effect on the profit of big brand with a popular fan page (L De Vries, 2012). Similar to the benefit of Yelp platform in the food industry, there is a huge impact of social media in the field of travel and hotel booking(Z Xiang, 2010).

In the field of ensemble training, good result for multi-label classification using ensembles training has been recorded (J Read, 2008). Graph Convolutional Neural Networks and the success of the model on the node classification task on Cora and Pubmed datasets has been introduced along with the se supervised approach to localized first-order approximation of spectral graph convolutions. (TN Kipf,2016). In this word, the aggregation is defined by two hidden layers with Relu activations followed by a Softmax function as in Figure 2. However, there are different ways to build the convolutional graph. By aggregate information from different numbers of step away from a given node instead of using its neighbor, we can have different results to fed to the network (Hamilton,2017). Method using GPU resources efficiently for computation on large graphs with billions of nodes has been proposed with the Pinterest dataset (R Ying,2018).



$$Z = f(X,A) = softmax(\hat{A}\ Relu(\hat{A}XW^{(0)})W^{(1)})$$

Figure 1. Multilayer Graph Convolutional

Figure 2. The definition of GCN graph

3.Datasets

3.1 Business dataset

This dataset has192.000 business and 20 features, which includes the average rating and user_id, which will be used to map a restaurant to customer. Most of the restaurant are in US, especially in Arizona and Nevada, and also many businesses in Toronto appeared on Yelp. Among 20 features, there are 19 features are categorical types, such as alcohol, parking or reservation. There are three attributes of the location of each business: postal code, city and state. When checked the distribution of ratings in each city, the result show that there is a similar pattern when the 3.5 rating always has the highest percentage. From this dataset, I will only remain the restaurant has more than 1 review and generate new value which is the mean and variance of neighbors of a restaurant in similar state, city and postal code. For example, will have restaurant k in the set of N restaurant, we will make the new attributes by calculating mean and variance of N-1 restaurant excluded k.

Mean: mean= $\sum_{i=0,i\neq k}^{N} \frac{1}{N-1} Xi$ Variance:$Var = \frac{1}{N-1} \sum_{i=0,i\neq k}^{N} (Xi - mean)^2$

Perform analysis on the variance of restaurant in same postal code, I see that same restaurants in a postal code could have very close rating because the value of variance is small. This can be explained by the fact that that neighbors can affect the success of a business by many aspects (L Hu, 2014). Characteristics of closer neighbors can affect the rating of business since users when vistit a business can also vistit its' neighbors. The closer the neighbors are, the more chances their characteristics can affect the rating. Compare to the statistics of the dataset, variance of restaurant by state is much higher than variance of restaurant by city and by postal code.

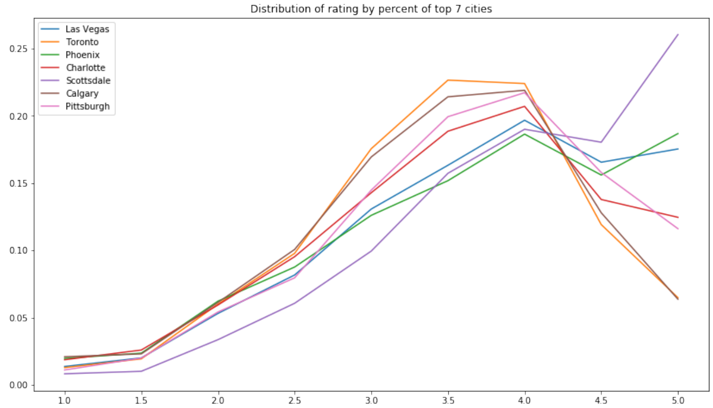


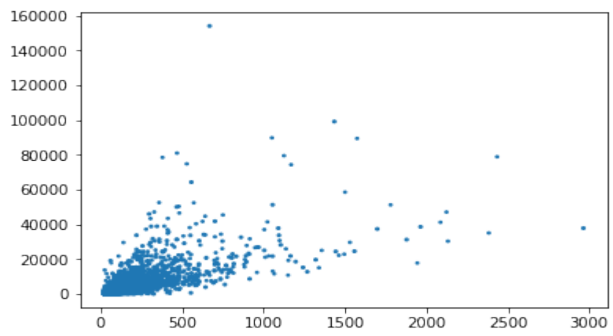
Figure 3. The distribution of rating of top 7 cities

Average variance by postal code	0.00032
Average variance by city	0.00067
Average variance by state	0.00113

Table1. Statistic of the variance

3.2 Users dataset and review dataset.

The user dataset contains information of ~ 1.6 million users which 18 attributes while the review dataset has ~ 6.8 million reviews and 8 attributes. In the user dataset, we have 4 features to show the characteristics of a user: useful, funny, cool, fans and besides these four is the characteristic of their compliment. For the pre-processing step I will only remain the users who gave more than 1 review. There is a correlation between number of fans and number of useful reviews, which suggest that user giving more useful reviews could have more fans. In the user dataset, the percent of average rating that users gave to restaurants with values of 3 and 4 are highest. One more thing to notice is that there are nearly 10% of total users have 0 fans, and on social media, the affection of a user can be affected by number of fans/followers (M Cha, 3220). In this case, user with 0 fan will have less effect on the rating of restaurant than user with 100 or more fans. In the review dataset, we have the mapping between a user_id and a a business_id. A user is mapped to a business if and only if the user has at least one review for the restaurant. Also, in the review dataset, we have the textual information of the review, however in this project I will not use this information. This textual information can be used for NLP task for extracting most important things that business should focus on to satisfy their customers.



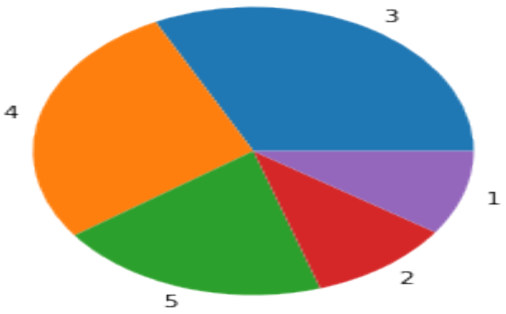


Figure 4. The correlation of fans and useful Reviews

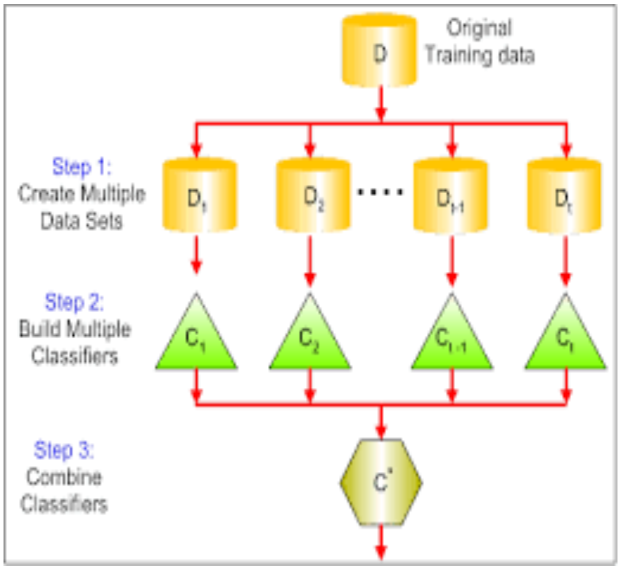
Figure 5. Percent of average rating score given by users

4 Methodology

4.1 Ensemble Training and classifying 9 classes

Ensemble method has been well-known for the ability to improve the accuracy and performance of predicting task in many fields of research (L Rokach, 2010). Severity of road traffic accidents has been classified well with the appliance of ensemble training (SY Sonh, 2003). Other real-word applications also had been tried with ensemble methods (Nikunji C.Oza, 2007). In my work, I decided to use two ensemble methods: Voting Classifier and Bagging. These models will be trained and tested on the business dataset. There are 9 classes to be classified {1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5} and all features exclude mean and variance attributes in the dataset will be converted to categorial features. For the Voting Classifier, I will use 4 base models as input for the voting model and take the output of the voting model as the output result. The four chosen models are based on four different machine learning algorithms : K Nearest Neighbors algorithm, a method given fast prediction based on the k number of most similar example; Decision Tree classifier algorithm which is good with using questions related to features to classify the input; Naïve Bayes used probability to predict the outcome and the last one, Support Vector Machine (SVC). The four based models will have different weights rated by the accuracy of each based model. The voting classifier will optimal the set of different classifiers and combined is by a specific fusion equation to give the output (Dymitr Ruta, 2004).

The other ensemble training is Bagging algorithm, which used many random sub examples and trained them on different decision tree model to boost the accuracy. This method will help reduce the variance of the model and given better results. (P Buhlman,2002). Using these two algorithms, I hope with the limit and one side information only from the restaurant dataset, I still can have a good prediction.



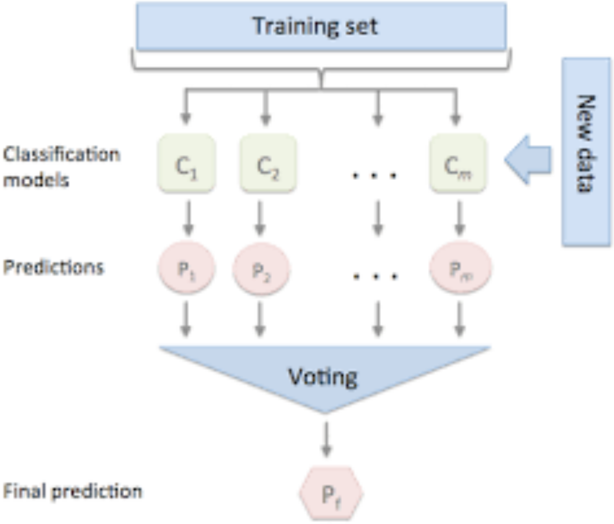


Figure 6. Bagging classifier model

Figure 7. Voting classifier

4.2 Graph Convolutional Neural Network

4.2.1 Graph definition and Node features

The network between users and restaurants will be a bipartite graph, where we have a set of users and a set of business. A link from a user to a business will be added if the user gave a restaurant a review and a rating score more than 3.5. In order to reduce the number of links, I only connect a user giving a restaurant a score more than 3.5 with a link, and when a user give the restaurant score less than 3.5, the link will not exist. Also, because we are trying to target the group of potential customers, I assumed that we should focus on users who like the restaurant and give the restaurant a high rating score. All link will be stored as an adjacency matrix.

$$G = (V, E) \qquad V = \{a \in Users, b \in Businesses\} \qquad E = \{(a, b)\ if\ a\ gave\ b\ score > 3.5\}$$

Each node in the graph, business or user, will contain the same information as in the businesses data set and the user dataset. As in the definition of graph, node a has all 18 features and node b has 26 features. All data points are normalized and pre-processed as in the Ensemble training part.

4.2.2 Graph model

Graph convolutional neural networks have been proved with the good results on link prediction and node classifications tasks (Kipf, 2016). One advantage of GCN model is that we can share many parameters at the same time, compare to normal approach often train one unique embedding vector for each node lead to the linear relation between number of nodes in the graph and number of parameters grows (A Grover,2016). GCN model can require retraining of model when every new nodes is added to the network, which different from the method of generate embedding for only training nodes (B Perozzi,2014). GCN and the inductive method are suitable for maintaining the dynamic change of social media platforms, in this case is Yelp.

The GCN model are used to combine the information about a business's local neighborhood and the embedding of the business itself, then use the result for the neural network in order to give the output predicting the star rate of the business. For each business, we are combining the information from the business, the profile of users who give the business good rating. I choose GraphSAGE (W Halmilton,2017) prototype to build a 1- layer graph convolutional neural network and take the input follow by this formula:

$$x_v^k = Relu\left(W_k\left[\frac{\sum_{u\in N(v)} X_u^{k-1}}{N(v)}, X_v^{k-1}\right]\right) \; with \; k > 0$$

Form this formula, the output of our model is the weight matrix W_k to apply for any node in order to get the low dimension embedding. I will train the GCN model 50 epochs, with the train: test ration of 0.3.

5 Results

5.1 Ensemble training models

The K-Nearest Neighbors has the lowest accuracy since it faced with the struggle of predicting 9 labels, while dimension is high and there may not have enough neighbors to compare with. The same struggle with the high dimension appears on SVC model. Decision Tree models and Naïve Bayes models has worked pretty well and given a decent prediction on the problems. The model based on Bagging algorithm gave better results than the four previous models, since it had combined different decision tree models to produce the output. The last model, voting classifier, is the one with the highest accuracy. The weight for based models is [1,1,2,2] based on the rank accuracy of based models. By combining different models with different approach and advantage, the voting classifier have returned a very good prediction despite the fact of missing data from the aspect of users. It proved that we can still produce good classifier based on the data about restaurants.

Model name	Accuracy
kNN	44%
SVC	52%
Decision Tree	63%
Naïve Bayes	63%
Bagging	71%
Voting	80%

Table 2. Prediction accuracy of different models

5.2 Graph Convolutional Network

The Graph Convolutional Network model trained on the network of users and business, with 50 epochs, achieve accuracy of 88% - higher than the previous Ensemble training models. This result show that by use much more information from the user dataset, we can increase the predicting efficiency is a notable scale. Also, this also show that profiles of different users can affect the rating of business which served the users in different ways.

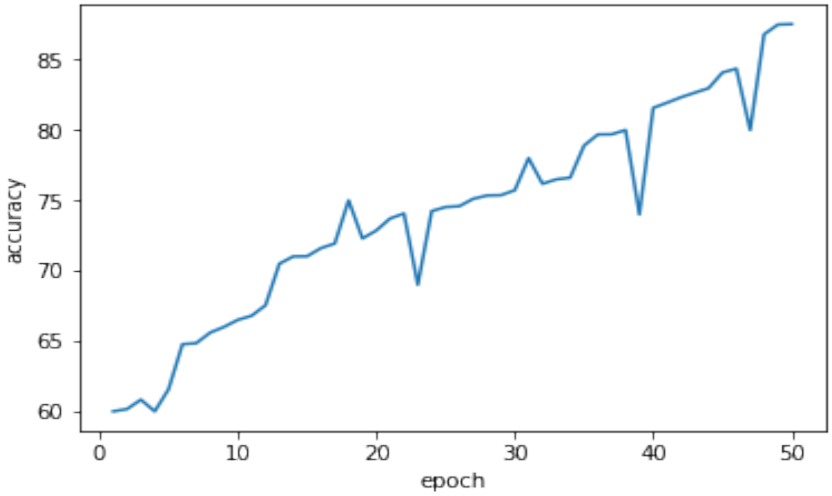


Figure 8. The accuracy rate of GCN model recorded by epoch

5 Conclusion

Comparing the results from two methods, we can see that using Graph Convolutional Network with the network of users and business will give us a better prediction than using only dataset of businesses and ensemble training model. Because when we use the network, we have bigger dataset than using only the business model, and the Graph Convolutional Network give us a good aggregation of node, as well as an efficiency embedding node, we have better result with Graph Convolutional Network.

6 Discussion

Although we have good results on predicting the star score of businesses on Yelp, there are still several point we have to notice. First of all, most of the restaurant on this data is coming from the West side of US, such as Arizona or Nevada, while there are a very little percentage of restaurants in D.C or New York, so when apply the model to predict the star score of restaurant in the East coast, we can have over-fitting problems and the accuracy will not be that high.

In addition, there are some way that we can improve the accuracy the model. First of all, an edge only be added if the rating of user to business is higher than 3.5, so there is a decent amount of data loss. If we create a weighted graph where each edge is actually the rating of a user to a business, we can reduce the loss of data, but also increase the training complexity because of the increased of network size. Moreover, I only trained the model with 50 epochs because lacking good GPU, so if we can train the model with bigger data and large amount of time, we can still improve the accuracy score of the model.

In the review dataset, there are a vector of the textual content of the review, such as "This food is so amazing!". Because the linguistic and language use can express the both the satisfaction and unhappy, we can perform sentimental analysis on these textual contents in order to improve the performance (J Rowley Maryfield,1998). For example, for each review, we can use sentimental analysis technique to extract the sentiment score of the review or for each users, we can take all of his or her review, and use a bag of words method to build a categories and topic of trending words that this user use (HM Wallach, 2006). Also, we can predict the review of a restaurant based on the information of restaurants in business data and the textual content of reviews in the review dataset by predicting the probability of certain words will be used for the restaurants

7 Acknowledgements

Thanks to Professor Kipf and his team for the work on Graph Convolutional Network and the instruction to build an efficiency GCN model.

Thanks to Professor Longke Hu and his team for the research on the effect of geographical neighborhood to the rating and reputation of businesses.

8 References

Aula, P. (2010). Social media, reputation risk and ambient publicity management. *Strategy & Leadership*, *38*(6), 43-49.

Qualman, E. (2010). *Socialnomics: How social media transforms the way we live and do business*. John Wiley & Sons.

Peter, J. P., Olson, J. C., & Grunert, K. G. (1999). *Consumer behaviour and marketing strategy* (pp. 329-48). London: McGraw-Hill.

Law, J. (2004). *After method: Mess in social science research*. Routledge.

Levinson, P. (2009). *New new media*. Boston: Allyn & Bacon.

Martilla, J. A., & James, J. C. (1977). Importance-performance analysis. *Journal of marketing*, *41*(1), 77-79.

Luca, M. (2016). Reviews, reputation, and revenue: The case of Yelp. com. *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, (12-016).

Huang, J., Rogers, S., & Joo, E. (2014). Improving restaurants by extracting subtopics from yelp reviews. *iConference 2014 (Social Media Expo)*.

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems* (pp. 3844-3852).

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 51-60).

De Vries, L., Gensler, S., & Leeflang, P. S. (2012). Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of interactive marketing*, *26*(2), 83-91.

Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism management*, *31*(2), 179-188.

Culnan, M. J., McHugh, P. J., & Zubillaga, J. I. (2010). How large US companies can use Twitter and other social media to gain business value. *MIS Quarterly Executive*, *9*(4).

Read, J., Pfahringer, B., & Holmes, G. (2008). Multi-label classification using ensembles of pruned sets. In *8th IEEE international conference on data mining* (pp. 995-1000). IEEE.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems* (pp. 1024-1034).

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, *53*(1), 59-68.

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., & Leskovec, J. (2018, July). Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 974-983). ACM.

Shirky, C. (2011). The political power of social media: Technology, the public sphere, and political change. *Foreign affairs*, 28-41.

Hu, L., Sun, A., & Liu, Y. (2014, July). Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 345-354). ACM.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010, May). Measuring user influence in twitter: The million follower fallacy. In *fourth international AAAI conference on weblogs and social media*.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1-2), 1-39.

Sohn, S. Y., & Lee, S. H. (2003). Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. *Safety Science*, *41*(1), 1-14.

Oza, N. C., & Tumer, K. (2008). Classifier ensembles: Select real-world applications. *Information Fusion*, *9*(1), 4-20.

Liao, Y., & Vemuri, V. R. (2002). Use of k-nearest neighbor classifier for intrusion detection. *Computers & security*, *21*(5), 439-448.

Ruta, D., & Gabrys, B. (2005). Classifier selection for majority voting. *Information fusion*, *6*(1), 63-81.

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, *30*(4), 927-961.

Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 855-864). ACM.

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014, August). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 701-710). ACM.

Rowley Mayfield, J., Mayfield, M. R., & Kopf, J. (1998). The effects of leader motivating language on subordinate performance and satisfaction. *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management*, *37*(3-4), 235-248.

Wallach, H. M. (2006, June). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning* (pp. 977-984). ACM.