Classification of News Headlines
An Annotated Bibliography

Arjun C (s4463233@student.uq.edu.au)
The University of Queensland

March 21, 2017

References

[1] E. Alpaydin, *Introduction to Machine Learning*, ser. Adaptive Computation and Machine Learning. Cambridge, MA: The MIT Press, 2014, vol. Third edition.

> In this book, the author Alpaydin has covered a wide range of topics required for implementing machine learning algorithms targeting different problem sets. The book not only includes the methodologies, but also the basics of topics like statistics and probability required to achieve a thorough understanding of how the algorithms work and should be built. The book covers many variations of machine learning algorithms and architectures like supervised learning, Bayesian decision theory, parametric methods, multivariate methods, multilayer perceptrons, local models, hidden Markov models. It also discusses the methods to assess the performance of the algorithms and to compare the algorithms. This book will help my project as it serves as the basis of understanding the topics/algorithms that I would be using in the project. This would help me in building multiple algorithms, compare their performances and to infer which performs better for the classification problem set that I would be working on.

[2] I. Dilrukshi, K. D. Zoysa, and A. Caldera, "Twitter news classification using svm," in *2013 8th International Conference on Computer Science & Education*, Conference Proceedings, pp. 287–291.

In this paper, Dilrukshi et al. discuss about using Support Vector Machines to solve the classification problem set retrieved from twitter and by selecting the required features on the basis of certain criteria. The authors use the data retrieved from tweets by the News groups and eliminating the words with highest and least frequency. Bag-of-words approach is used to carry out the further process. They make use of SVM to classify the news into 12 pre decided groups. 90% of their data was used for training and 10% for testing. The average accuracy achieved was 75%. The authors base their selection of the classification model on previous research which may not hold good for this problem set. Even though the approach of eliminating the words seems like we may be leaving out some of the required features, it does shed some light on a possible approach to effectively tackle the given problem. This paper will help me understand how to go about tackling the problem of considering the words in continuous text rather than just the numerical values.

[3] S. O. Fageeri, S. M. M. Ahmed, S. A. Almubarak, and A. A. Mu'azu, "Eye refractive error classification using machine learning techniques," in *2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*, Conference Proceedings, pp. 1–6.

In this paper, Fageeri et al. discuss about the classification of eye disorders using SVM, naive Bayesian and J48 decision trees applied on a real dataset . They discuss about the possibilities of detecting disorders in the eye in an earlier stage based on the recorded samples of individuals which would help in overcoming any possible visual impairments. The axis, sphere and cylinder of both the eyes being the main features in this classification as discussed in the paper, they do mention data preprocessing using Weka. The removal of noisy data and the missing data are taken care of by Weka. The key processes to consider for my project here would be the comparison between the three classifiers used and the accuracy of them for this particular dataset. A collective high of the accuracy of one of these algorithms could be one of the aspects to consider for any classification datasets. This helps

my project by providing me an intuition that all algorithms need not work the same for all the datasets.

[4] F. Figueiredo, L. Rocha, T. Couto, T. Salles, M. A. Gonalves, and W. Meira Jr, "Word co-occurrence features for text classification," *Information Systems*, vol. 36, no. 5, pp. 843–858, 2011.

In this article, Figueiredo et al. have brilliantly put forward an approach for text classification which uses word co-occurrence to reduce the noise in Bag of words approach. They have discussed the Syntactic phrases approach, Word n-grams sequence approach as well. They suggest ways to increase accuracy by feature selection and feature pre-processing. The classification was carried out using SVM, kNN and Naive Bayesian methods, SVM performing the best for this dataset and an increase in the accuracy with the suggested approach is seen. The approach and the way dataset is considered appears to be more feasible as it makes classification less prone to error for a document with complete content than just the headline or the abstract. This article helps my project in understanding the ways to reduce the noise in Bag of Words approach, the importance of the data we consider for the classification and feature preprocessing.

[5] S. Hans Georg, *Support Vector Machines*. Wiley-IEEE Press, 2012, pp. 179–196. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6331082

In this extracted section, the author Hans gives a complete picture of Support Vector Machines. He discusses about using SVM for a linearly separable problem and the mathematical solution on how to achieve the requirement. He then depicts how, many problems are not linearly separable and the issue with using SVM in these situations. He then discusses about the Kernel function using which we could solve non-linear classification problems, also targeting the multi class classifications. This section gives an in-depth understanding of Support Vector Machines and possible hurdles that we need to be aware of before building the algorithm. Using the concepts, we could restrict the implementation of the algorithm to target specific problem set. This section helps my project

in implementing the algorithm in addition to the concepts explained in "Introduction to Machine learning" by Alpaydin.

[6] R. Khanna and M. Awad, *Efficient Learning Machines Theories, Concepts, and Applications for Engineers and System Designers*, ser. The expert's voice in machine learning. Berkeley, CA : Apress : Imprint: Apress, 2015.

In this book, the authors Rahul and Mariette bring out a very practical approach for the users to learn machine learning. The readers are introduced to the different algorithms based on appropriate grouping and what they could be used for. In addition, the case studies for the prominent algorithms and a set of case studies to tackle real life scenarios give readers a better understanding of the applications of the discussed learning methods. The visualizations provided help the readers in picturing some of the aspects that are sometimes just explained in code or mathematical perspective. The properties of SVM and a comparison of ANN and SVM was worth noting. This book covers classification topics like Logistic Regression, Naive Bayes and SVM which are of interest for the project process. This book helps me in getting a practical idea of the application of algorithms to the problem dataset.

[7] H. Kim and M. Kim, "Model-induced term-weighting schemes for text classification," *The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, vol. 45, no. 1, pp. 30–43, 2016.

In this article, the authors Hyun and Minyoung suggest and review the improvement in the performance of a classifier by implementing term-frequency based weighting than just the term-frequency approach for text classification. The idea discussed here is that some words have a very high possibility of appearing in text of a single class (e.g. 'Cricket' mainly appears in the sports category of news). By giving a weightage and not just the frequency to such terms, the probability of that particular text classified to a more appropriate class would be higher and this could help us in classifying the problem set with more accuracy. From the graphs we see that the results are promising with this approach implemented on mul-

tiple datasets than the existing ones. This helps my project in understanding the weighting approach that could help classify the problem set's data more accurately.

[8] Y. Ma and G. Guo, *Support Vector Machines Applications.* Cham : Springer International Publishing, 2014.

In this book, the authors Yunqian and Guodong have brought together works of many other researchers on SVM, its variations and their applications. One of the chapters (Chapter 2) focuses on multi class classification using SVM. In addition to the approaches to solve multi class problem, their advantages and disadvantages in terms of complexity, performance and computational requirement considering a specific scenario are also discussed. The experimental evaluation of the approach SimMSVM as depicted in the tables suggests an increase in the training speed and the accuracy of the prediction. This lets the readers think of possible ways to improve the performance of the process as a whole. This chapter from the book gives an extended understanding of SVM to target the problem I should be working on. Also, another chapter (Chapter 5) discusses the Bag-of-features model for image classification where the similarity with the Bag-of-words approach is brought up with the implementation comparison which gives a comparative understanding of the approach.

[9] Z. Niu, Z. Yin, and X. Kong, "Sentiment classification for microblog by machine learning," in *2012 Fourth International Conference on Computational and Information Sciences*, Conference Proceedings, pp. 286–289.

In this paper, the authors Zhen Niu et al. discuss a step by step approach for sentiment classification. They suggest the steps to be carried out for textual analysis listed as extracting feature items, calculating the weights, training samples, sentiment classification and performance evaluation. They talk about three filtration steps for removing the words that are most likely going to be noise for training and give an idea on what words to keep. They calculate the weights by comparing the possibility of known classifications with conditional probability of classification after giving the weights. Naive

Bayes is used to train the model. Even though the probabilistic approach seems more feasible, the article only says that the efficiency is high but there isn't any visualization or numerical proof suggesting the same. This article helps me in understanding one of the ways to select the features, determine weights and the evaluation method for my classification problem set.

[10] P. Tsangaratos and I. Ilia, "Comparison of a logistic regression and naive bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size," *Catena*, vol. 145, pp. 164–179, 2016.

In this article, the authors Tsangaratos and Ilia discuss about comparing logistic regression and Naive Bayes for multiple combinations of dataset size and the model complexity. The importance of evaluating such a combination is made clear from the experimental results. The process of preprocessing the data with the underlying processes and the need for these steps are made clear. Training, prediction and evaluation steps are given a brief explanation. This article focuses on LR as it is considered a very good approach for this problem set type. A comparison with algorithms like SVM or Random Forests would have given a better picture. Nevertheless, the simplicity of logistic regression gives scope for easy understanding of the behavior seen in the classification accuracy and the performance. This article shows the importance of such evaluations for any classifier and the respective dataset thus providing me an additional task to consider for such validations.

[11] A. A. Turdjai and K. Mutijarsa, "Simulation of marketplace customer satisfaction analysis based on machine learning algorithms," in *2016 International Seminar on Application for Technology of Information and Communication (ISemantic)*, Conference Proceedings, pp. 157–162.

In this paper, the authors Ajeng and Kusprasapta discuss their idea on sentiment analysis based on twitter data using five learning methods to select the better performing algorithm. The procedure followed is explained in very simple steps. The preprocessing of data is carried out in five sequen-

tial steps i.e. case folding, tokenizing, filtering, stop word removal and stemming thereby removing noise that could degrade the performance. The use of TF-IDF weighting process is mentioned which increases the probability of being classified into the right groups. The training process is then carried out using five different classifiers and their accuracy is calculated. This paper helps me in understanding how the data has to be preprocessed before using the Bag of words approach to go ahead with the training of the model. The tests by the authors have shown that SVM was a better performer for the selected dataset compared to the other classifiers. This outcome along with the outcomes considered from other references suggest the use of SVM for the classification dataset.

[12] N. Wang, B. Varghese, and P. D. Donnelly, "A machine learning analysis of twitter sentiment to the sandy hook shootings," in *2016 IEEE 12th International Conference on e-Science (e-Science)*, Conference Proceedings, pp. 303–312.

In this paper, the authors Nan Wang et al. share their findings and the issues they came across when they performed the sentiment analysis on twitter data. They have used the approach of splitting the text into unigrams, bigrams and trigrams and using Hashtags as additional features as they depict a strong classification factor. They have used eight predictive models to compare performances for multiple combinations of sample data count and n-gram value. An interesting factor is the inclusion of population for the normalization to provide a correct picture of the pro-gun sentiment. A mention of Part-of-Speech (POS) tags not being a strong indicator of emotions was made. The explanation on how using dictionary approach to classify the text based on frequency is not a good idea is made clear. This paper makes me consider the n-gram approach and gives a better understanding of data preprocessing to attain better results for classification problem dataset.