## **Brief History on Natural Language Processing and Evaluating the Effectiveness of Text Classifiers Created Using Machine Learning**

G.Arunath, S.M.V Dilshan

British College OF Applied Studies, Kandy, Sri Lanka

[arun@bcas.lk](mailto:arun@bcas.lk), [smv.dilshan@gmail.com](mailto:smv.dilshan@gmail.com)

Natural Language Processing (NLP) helps computers understand and use Human Language. Text Classification is a Subfield in NLP that deals with categorizing given text into organized groups. In Machine Learning computer systems are created that can learn to achieve tasks without explicitly being programmed. Given the Vast Amount of Unstructured text data that is created today text classification and NLP should play a vital part in analyzing them. This will help in reducing the workload of humans in analyzing these data. The aim of this paper is to give the reader an introduction into NLP and Text Classification theoretically and show the effectiveness of each of these chosen algorithms in different classification tasks. This research will help other researchers in selecting the best algorithm for a given text classification tasks and push the field in new directions.

**Keywords:**

Natural Language Processing, Machine Learning, Statistical Learning, Text Classification

# **Introduction**

In this information Age the Amount of data in unstructured format is staggering. If someone were to shift through this data, it will be exhausting and inefficient. Creating a computer Program That can understand language will be a very complex task that need domain knowledge from many different Fields like Computer Science, linguistics and Mathematics. But we can use Machine Learning Methods and let the software model find out the relationship between the words by itself given supervision. This task will seem daunting from Beginners Perspective, so this paper will provide simple explanations on how these algorithms work, what is natural language processing, how to choose the best algorithm to a given task. Main Focus Will be to Give the Reader an Understanding on What is Natural Language Processing and differences in the Algorithms used for Text Classification

# **Methodology**

This paper will use Both Primary and secondary research for finding facts. First task was searching for all the Papers, Journal Articles Written About Natural Language Processing and History of Natural Language Processing, choosing from them what to extract given research context. Then Creating a Timeline for the advancement of Natural Language Processing from its inception from the Facts that were extracted.

The Second Phase was finding what are the main Approaches for NLP (Statistical and Rule Based) and finding statistical algorithms that can be used for text classification using Secondary research. During this Stage the following algorithms was chosen to evaluate on K- Nearest Neighbors, Native Bayesian, Support Vector Machines, Logistic Regression and Random Forest.

Final Part of the Research was evaluating these Approaches. The Algorithms ran on 5 different datasets with different classifications to get higher accuracy and the accuracy was measured using a confusion matrix and a classification report. Both are accuracy measuring methods in sikitlearn Python library. Researchers evaluated how each algorithm performed on each of the given datasets using each of the algorithm’s accuracy, consistency, and lack of false negatives on all given datasets.

The Algorithms where implemented using python programming language and Jupiter notebook Software. External libraries Used: Pands (data manipulation), scikitLearn (algorithms, accuracy methods), numpy (mathematical functions).

# **Experimental Results**

Natural Language Processing has been associated with Artificial Intelligence since its conception in 1950’s. It evolved alongside Artificial Intelligence going through similar advancements and pitfalls along the way. When natural Language processing was starting most of the algorithms used were hand coded. That is, all the rules of the language should be defined and coded. This may sound achievable, but each language is very complex and the rules, conditions that needed to create these models/ algorithms were staggering. When the Machine learning algorithms were introduced into NLP the whole approach to the problem were shifted to creating and optimizing models that could learn these complex rules of language. These models outperformed most of the classical algorithms that used rule-based methods. Another advantage of this approach was unlike the Rule Based Model the Engineers didn’t need a background in linguistics to make an NLP model. The main problem with the Machine Learning approach was it needed a sufficiently large dataset with labeled or unlabeled data to make the most out of this approach.

When each of the Dataset was classified using the given Classification models it can be observed that by comparing between Spam Email dataset and Spam messages dataset Random forest and K nearest algorithms works well when given two categories to categorize to and furthermore it can be observed that K nearest algorithm works well when it has more data to work with (spam text has less words 30 or less words per each message).

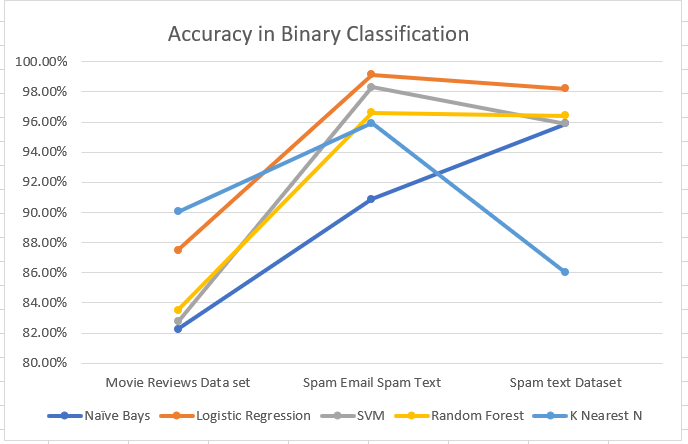


Figure Average Accuracy in Binary Classification

When taking the accuracy average across all the tests logistic regression seems to perform well given different datasets with different attributes (Smaller word count, more categories). When selecting these models for classification if the dataset has 2 categories the users can use one from Naive Bays, Logistic Regression, Support Vector Machine, but here if each instance of data contains less words (30 or less) using Naïve bays or k nearest classifiers is ill advised because it’s the worst performing algorithm when there are less words to work with.

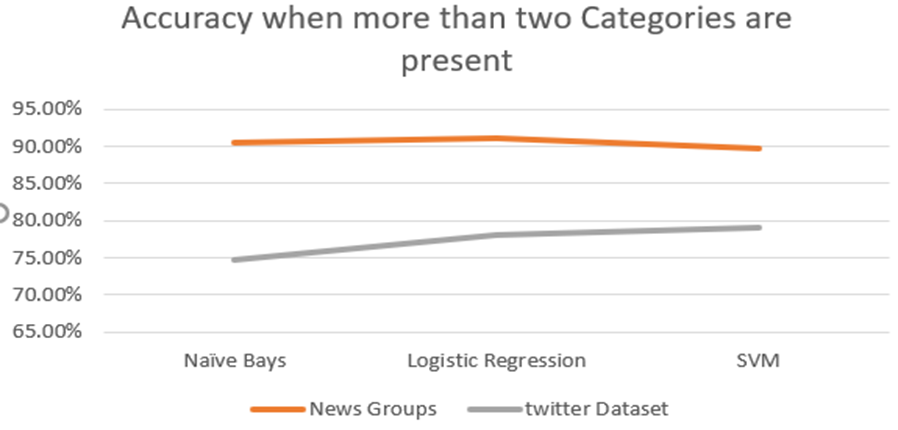


Figure Accuracy Given Many Categories

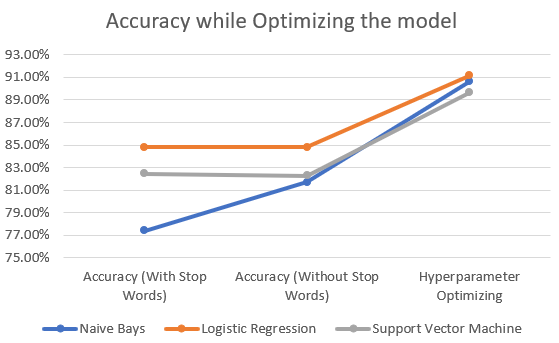


Figure Optimizing the Model

If the dataset contains more than 2 categories (in this example 20) best selection would be Logistic regression classifier coming in top with the highest accuracies followed by support vector machine and Naïve bays. Naïve Bays and Logistic regression both improved due to a cleaner dataset. When using these classifiers cleaning the data beforehand is well advised. By using the above results, it can also be observed that the word count (removing stop words) does not affect the support vector machines, logistic Regression and the Naïve Bays models.

The dataset can be cleaned by removing unnecessary symbols from data and by removing stop words (common occurring words that have no impact on the meaning of the sentence). Then tuning some of the hyper parameters that is used in the classifier can produce a model with higher accuracy.

# Conclusion

When Machine Learning is applied to Natural Language Processing it creates models with higher accuracy without needing to code each complex rules of language in to creating the model. These models can be trained to do any type of text classification task given a clean dataset and a good training cycle. This is a far cry from the Rule Based approach that uses classical algorithms. The researchers only needed to have a basic understanding of linguistics to create these models. These classifiers this can be very helpful when organizing/classifying large amount of electronic data. From the above data it can be surmised we can use logistic regression to get a highly accurate model that can classify text with Naïve Bays and Support Vector MacInnes closely following in with similar accuracies given any dataset. Given machine learning Algorithms performs differently according to the nature of the training dataset, it is highly advised to perform data preprocessing before using the data to train the Machine Learning Model.

It can be clearly stated that Using Machine Learning Methods for Text Classification is far more efficient than using the classical way. Using this method any person motivated, intuitive enough can create a model that can classify very complex text documents.

# References

Raval, S. (2019). Natural Language Processing. [video] Available at: https://www.youtube.com/watch?v=bDxFvr1gpSU&t=6s [Accessed 5 May 2019].

Janakiev, N. (2018). Practical Text Classification With Python and Keras – Real Python. [online] Realpython.com. Available at: https://realpython.com/python-keras-text-classification/ [Accessed 17 Jun. 2019].

James, G., Hastie, T., Witten, D. and Tibshirani, R. (2017). An introduction to statistical learning. 7th ed. New York: Springer.

Zafra, M. (2016). Text Classification in Python. [online] Medium. Available at: https://towardsdatascience.com/text-classification-in-python-dd95d264c802 [Accessed 5 May 2019].

Korde, V., & Mahender, N. C. (2012). Text classification and classifiers: a survey. International Journal of Artificial Intelligence & Applications, 3(2), 85-99. https://doi.org/10.5121/ijaia.2012.3208 [Accessed 6 May 2019].

Nadkarni, P., Ohno-Machado, L. and Chapman, W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, [online] 18(5), pp.544-551. Available at: https://doi.org/10.1136/amiajnl-2011-000464 [Accessed 16 Apr. 2019].