# Literature Survey

The article written by Dr. Sebastian [1] describes how Naïve Bayes Algorithm can be utilized to classify a given email message as spam or ham. Bag of Word model is used which was constructed after pre-processing an email message through stop-word removal, stemming and lemmatizing. Natural Language Toolkit [2] is an open source program developed to perform various text processing tasks. This package defines methods for stop word removal, lemmatization and stemming. The article [3] is based on n-gram model for spam filtering where they found out that by using a n-gram of size 3 and 4 would produce better results when compared to the normal unigram model.

The web article [4] shows a step by step expressions and values for deciding if a sentence is sports or non-sports related. Here they implemented a count based approach and used this as prior for calculating the conditional probabilities. They also show how to make use of Laplace Smoothing to prevent probabilities to always be zero. Another interesting web article [5] which uses the concept of logarithms to prevent underflow while working with large denominators. On applying the logarithm for each term of the Bayes Theorem, the multiplication would be converted to addition and division would be a simple subtraction of the logarithms of each terms.

# References

|  |  |
| --- | --- |
| [1] | D. S. Raschka, "Naive Bayes and Text Classification," 4 October 2014. [Online]. Available: https://sebastianraschka.com/Articles/2014\_naive\_bayes\_1.html. |
| [2] | E. Loper and S. Bird, "NLTK: The Natural Language Toolkit," 17 May 2002. [Online]. Available: http://www.nltk.org/. |
| [3] | I. KANARIS, K. KANARIS, I. HOUVARDAS and E. STAMATATOS , "WORDS VS. CHARACTER N-GRAMS FOR ANTI-SPAM FILTERING," *International Journal on Artificial Intelligence Tools ,* pp. 1-20, 2006. |
| [4] | B. Stecanella, "A practical explanation of a Naive Bayes classifier," 25 May 2017. [Online]. Available: https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/. |
| [5] | H. H. Nguyen, "Algorithms for Text Classification," 4 February 2019. [Online]. Available: https://towardsdatascience.com/algorithms-for-text-classification-part-1-naive-bayes-3ff1d116fdd8. |