On the Naive Bayes Approach to Parts-of-Speech Tagging

Bart Gleen Q. Sanchez Rey R. Cuenca*

May 2022

Abstract

This paper serves as an exposition of the Naive Bayes approach to parts-of-speech (POS) tagging.

1 Introduction

Internet connectivity and digitization have ushered the era of information explosion specifically in the areas whose generation and consumption involved textual data written in the natural language. Consequently, the indispensability of machines in automating human tasks specifically dedicated for processing such large volumes of information has been increasing exponentially for the past two decades under the banner of the subfield intersecting computer science, linguistics, and artificial intelligence (AI) called Natural Language Processing (NLP).

To design algorithms that can enable machines to imitate the way how effortlessly humans process natural language is quite a herculean tasks, a fact that many computer scientist have unanimously admitted. This is because natural (human) language is fraught with both complexity in structure and ambiguity in meaning. NLP plays a significant role in addressing these issues inherent to natural languages.

One of the most important addressed topics and primary building blocks and applications of NLP is what we call the *parts-of-speech* (POS) *tagging*, also known as *grammatical tagging*. The process of POS tagging involves assigning each word (*token*) in a text the appropriate syntactic tag in its appearing context, e.g. sentence. A syntactic tag called *parts-of-speech* (POS) is a grammatical category that agreed upon by linguists to be possessed by a certain kind of natural language system. To name a few, some of these tags include the verb, adjective, adverb, and noun.

Machine translation, word sense disambiguation, question answering parsing, and other very important applications use POS tagging as preprocessing.

*Department of Mathematics and Statistics, MSU-Iligan Institute of Technology

The origins of POS tagging can be traced back to the ambiguity of many words in terms of their function in a given context.

Labeling each word in a sentence with its appropriate part of speech is easy for humans. When we encounter sentences, we assign POS tags based on their context. For example, in the sentence, "David has purchase a new laptop from apple store". In this sentence, every word is associated with a part of speech tag which defines their functions. DAVID has an NNP tag which means it is a proper noun. Further, HAS and PURCHASED belong to the verb indicating that they are the actions. The LAPTOP and APPLE STORES are the nouns. NEW is the adjective whose role is to modify the context of the laptop. However, human can't able to labeled everything manually if it has a huge data.

As a result, it is necessary to have pre-tagged data that has been tagged by human specialists in order to determine if the assigned tag is valid or not. This may not appear to be a tough work to a person, but the huge number of difficult terms found in natural languages makes this a challenging assignment for a computer. POS tagging is now done in the context of computational linguistics, using algorithms which associate discrete terms, as well as hidden parts of speech, by a set of descriptive tags.

The Bayes Theorem is its foundation. It is one of the most fundamental yet effective machine learning algorithms now in use, with applications. In this situation, the Naive Bayes classifier, would be the optimal choice because it selects the class with the highest conditional probability for a target tokens. Consequently, this model will be used to classify each tag in the sentence and determine the accuracy of the model's prediction. Due to its assumption of independence and superior performance in addressing multi-class problems, Naive Bayes is typically employed for text categorization.

For multi-class predictions if the assumption of feature independence holds, it can outperform other models while using far less training data. In Naive Bayes, it is assumed that all predictors (or attributes) are independent, which is rarely the case in practice. This reduces the algorithm's applicability in actual situations. Its probability results should not be taken seriously because its estimations are sometimes inaccurate. Naive Bayes is commonly employed for in-text classification due to its assumption of autonomy and strong performance in tackling multi-class problems.

This study is an exposition of the Naive-Bayes model as an approach to Parts-of-Speech tagging. We expound the theoretical underpinnings of the model as a graphical model and the details of it's advantages and disadvantages. Using the (insert the name of data set here) data set, we demonstrate how to perform POS tagging using the free statistical software R.

1.1 Statement of the Problem

The Naive Bayes classifiers are a class of probabilistic classifiers based on the Bayes theorem and a strong assumptions of feature independence. Despite its simplicity, it provides accurate text classification prediction. And it's been widely

employed in text classification in recent years. These algorithms are incredibly basic, quick and easy to understand and dependable.

This research is conducted to enhance the performance of naive Bayes model, which will be serve as the basis for the researchers for further research. Also, the study looked to update the profile of naive Bayes model who is said the best model in text classification and acknowledge the effectiveness in improving the model's prediction accuracy. This study specifically aims to determine the model's prediction accuracy by using the naive Bayes model.

1.2 Significance of the study

The findings will provide an information about the performance of naive Bayes model and the following are some of the advantages of using the Naive Bayesian classifier:

1. The model is less sensitive to outliers. When there is an outlier in the data, the findings can skew. As a result, data analysis necessitates the elimination of outlier effects. Because the Naive Bayesian classifier uses a probabilistic distribution, it can reduce the original data's outlier effects, making prediction results less sensitive to outliers.

2. The parameter can be predicted with fewer data. Relearning the probability of a given conditional probability is not strictly necessary in the Naive Bayesian classifier when utilizing the utility function, reducing data collecting efforts for risk prediction.

3. By decreasing the difficulty of data collection, the classifier is more applicable to industry practitioners. Despite the fact that the model relies on simple assumptions and procedures, the classification results are effective, even if the provided assumptions are incorrect.

4. The models allow for the incorporation of additional qualities by merging current attributes' characteristics, improving the model's prediction accuracy.

The researcher picked the Naive Bayesian method as the key mechanism for the proposed prediction model because of the aforementioned benefits. This study may serve also as a reference and guide for the future research studies.

1.3 Objective and Limitations of the Study

This study generally assesses in determining the model's prediction accuracy by using the naive Bayes model.

1.3.1 Objectives of the Study

Specifically, the study aims to:

1. To construct and investigate the performance of naive Bayes model in which able to select appropriate part of speech tag in the text.
2. To further improve the accuracy of POS tagging with proposed model, naive Bayes model.
3. Discuss the naive hypothesis and analyze the effect on classify performance, then present the corresponding improvements.
4. To enable the researchers to use the generated corpus for NLP applications.

1.3.2 Scope and Limitations

The present study looked to update the naive Bayes model in predicting the accuracy of the text. Also, determining the likelihood of a given sequence of words appearing in a text using naive Bayes model and its parameter to offer a word predictions to generate a text as an output. Moreover, the coverage of the study focus on the model itself which is the naive Bayes model and does not cover the other insignificant model in this study.

1.4 Definition of Terms

Polysemous -