PARTS OF SPEECH TAGGING USING NLP TECHNIQUE

Umme Athiya[1], Chaithra A S[2]

[1]*UG Student, Dept. of ISE, Don Bosco Institute of Technology, Bangalore, Karnataka*
[2]*Assistant Professor, Dept. of ISE, Don Bosco Institute of Technology, Bangalore, Karnataka*

---***---

Abstract - *POS tagging is a process of categorizing each word in a sentence into nouns, verbs, pronouns, adjectives, conjunctions and determinants. Usually systems employ this tagging process for preprocessing the text. There are mainly two approaches, they are Rule and Stochastic Based Approach. My proposed approach use lateral entries in the dictionary along with their information. Here, it mainly depends on the frequency of occurrence of words in the dictionary and identifying the meaning, the context they are referred as well.*

***Key Words*: POS, Stochastic, Rule Based, Hybrid Approach, Tokenization, Lemmatization, and Ambiguity Resolution.**

1. INTRODUCTION

POS which is also referred to as Parts of Speech Tagging is very useful process which is basically employed under Natural Language Processing concept such as to contextualize**,** information extraction and concepts like polarity and sentiments classifications etc.

2. BACKGROUND THEORY

In the background theory, it consists of two types of techniques, these are categorized as supervised and unsupervised learning systems. Supervised tagging is based on previously tagged corpus (or available dictionary tagged elements).

2.1 RULE BASED TAGGER

As the name itself specifies, the tagging system have rules assigned to determine the POS of words in a sentence. Although being employed with rules, this approach failed to tag some unknown words. Such type of systems developed failed due to unknown words. Hence certain good and efficient rules are needed to obtain a precise accuracy.

2.2 STOCHASTIC TAGGER

Stochastic approaches use pre trained and existing model to tag input text from the users. It is also a supervised learning method which builds a model based on tagged data (dataset). The words which are not encountered in the dataset trained it goes with the occurrence or probability, as these do not assign the correct tag since they are language independent. Most of the taggers are created based on models like Hidden Markov Model (HMM), Support Vector Machine (SVM), n-gram, Decision tree, Multinomial Naïve Bayes.

2.3 HYBRID APPROACH

Hybrid Approach is a phenomenon which uses the combined form of both the rule based and stochastic approach. Words in this approach first undergo the process of rule based approach where statistical rules are applied to the text. The next level is the stochastic based approach which goes with the trained dataset and thus check for the probability of occurrence of words in it.

3. PROBLEM STATEMENT

3.1 Existing System

Although there has been an enough advancement in the technology, still identifying POS Tagging is much more difficult when compared to matching words to their independent parts of speech via a dictionary method. It is quite different to sense that one word has two or more meanings based on their reference and the context they are being used. It is difficult to individually classify and assign the parts of speech for words manually. Thus new types of words keep building up in the dictionary, as POS depends on pre trained data it is not capable of scaling to newly introduced words.

3.2 Proposed System

The objective of this paper is to increase automaticity and maintain high precision, while limiting the size of human made corpus.

In our current work we approach the task of POS tagging as an optimization problem. Thereafter two new approaches based on the principles of single objective optimization and multi-objective optimization are proposed for POS tagging.

4. SYSTEM DESIGN

4.1Tokenization

Tokenization is process of dividing the stream of text entered by the user into individual group of words/ phrases/ symbols which give a meaningful elements. These process of division is the initial step which is further taken as the next level for the preprocessing. Tokenization is also referred to as text segmentation or lexical analysis.

4.2 Lemmatization

Lemmatization is the process of grouping the words together with the same meaning which was obtained from the tokenization process. Lemmatization is like the same way of stemming. So it connects words with similar sense with the other words. Text preprocessing combines both the process of stemming as well as lemmatization. Some treat these two as same. Actually, lemmatization is preferred over Stemming because lemmatization does morphological analysis of the words.

4.3 Removal of Stop Words

The Removal of stop words is the process where the most commonly occurring words in the document are extracted and using these words would mean to contribute very little meaning or contribute no help in the tagging system.

Stop Words

Stop Words are also referred as the "Bag of Words". These bag of words are commonly occurring for example 'a', 'an', 'the', etc words. These are explicitly recognized by the search engines and made to be ignored to contribute effective meaning in the understanding of the text or the document.

4.4 PARTS OF SPEECH TAGGING

The **Parts of Speech Tagging** is the process of classifying each word entered in the text or sentence into its corresponding tagging system such as Noun, Verbs, Pronouns, Determiners, Adjectives, Interjections, Conjunctions and many more.

4.5 AMBIGUITY RESOLUTION

Given a sentence with **ambiguous** words, its most likely to split the words into its lexical categories with various meanings for the same words based on the context and the information they are being referred in that particular situation.

4.6 HIGH LEVEL DESIGN

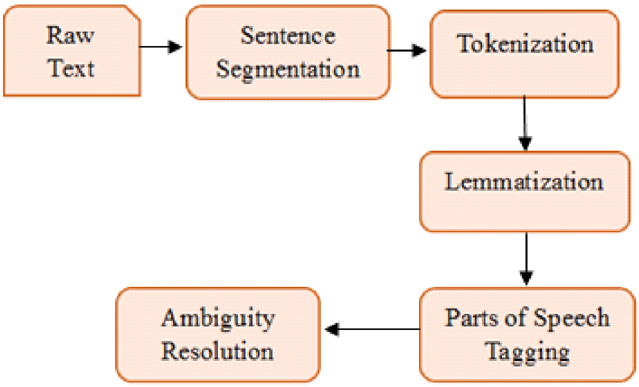


Fig 4.6.1:-Block Diagram for Parts of Speech Tagging.

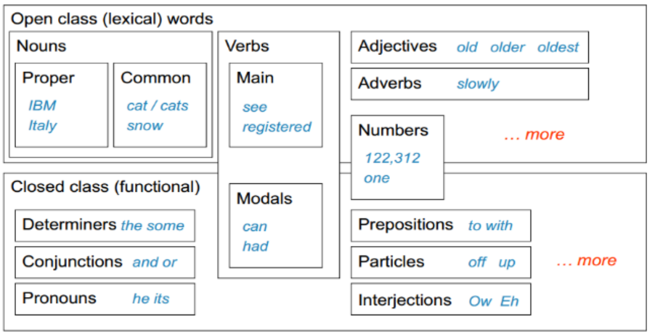


Fig 4.6.2:- Categories of POS

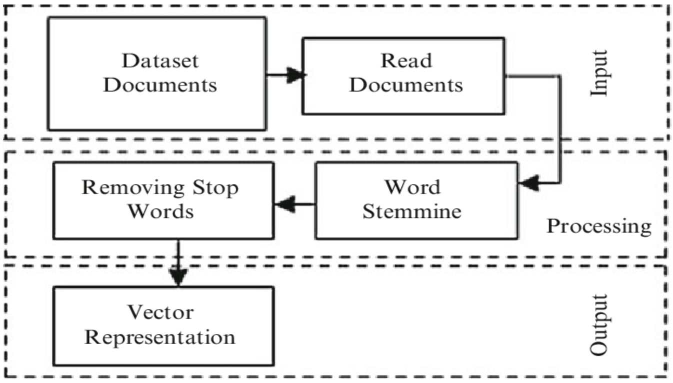
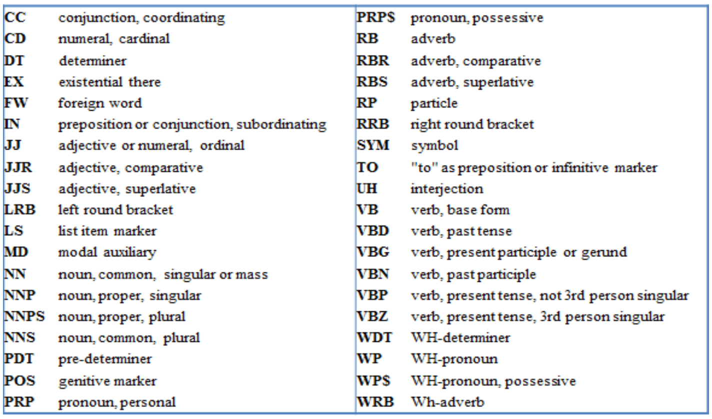


Fig 4.6.3:- Document Preprocessing Block Diagram

Fig 4.6.4:- Preprocessing



CC	conjunction, coordinating	PRP$	pronoun, possessive
CD	numeral, cardinal	RB	adverb
DT	determiner	RBR	adverb, comparative
EX	existential there	RBS	adverb, superlative
FW	foreign word	RP	particle
IN	preposition or conjunction, subordinating	RRB	right round bracket
JJ	adjective or numeral, ordinal	SYM	symbol
JJR	adjective, comparative	TO	"to" as preposition or infinitive marker
JJS	adjective, superlative	UH	interjection
LRB	left round bracket	VB	verb, base form
LS	list item marker	VBD	verb, past tense
MD	modal auxiliary	VBG	verb, present participle or gerund
NN	noun, common, singular or mass	VBN	verb, past participle
NNP	noun, proper, singular	VBP	verb, present tense, not 3rd person singular
NNPS	noun, proper, plural	VBZ	verb, present tense, 3rd person singular
NNS	noun, common, plural	WDT	WH-determiner
PDT	pre-determiner	WP	WH-pronoun
POS	genitive marker	WP$	WH-pronoun, possessive
PRP	pronoun, personal	WRB	Wh-adverb

Fig 4.6.5:- POS Notations

5. REQUIREMENTS

5.1 Hardware Requirements

Various features of a CPU that influence its speed power like bus speed, cache and MIPS are often ignored. The processor used is Intel core i5, speed is 1.1 GHz, and RAM is 8GB.

5.2 Software Requirements

Software requirements give a brief description of the software amenities that are required for the successful execution of the software with minimal errors. The software used for implementation are Jupyter notebook, Visual Studio Code for UI front end design.

6. APPLICATION

The dataset collected in here is the Women's Clothing E-Commerce Reviews which is used as Text Classification and Recommendation Classifier dataset. By entering the reviews for a particular product, the review undergoes the pre-processing and tokenization phases and thus by making analysis predicts whether the product is recommended or not been recommended.

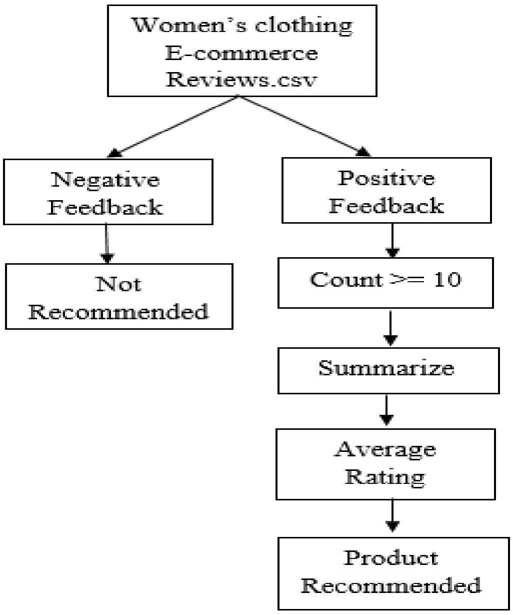


Fig 6: Use Case Diagram.

7. IMPLEMENTATION

7.1 Algorithm

7.1.1 Multinomial Naïve Bayes

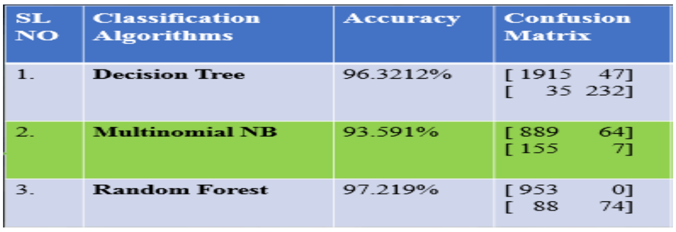
The Multinomial Naïve Bayes is a special instance of Naïve Bayes Classifier. A Naive Bayes classifier is a probabilistic Machine Learning model which classifies the instance based on the probabilities of occurrence of the words in the text or the document. Naïve Bayes Algorithm implements the feature or the concept from the Bayes theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Multinomial Naïve Bayes is mostly used for document classification problem, i.e whether a document belongs to the category of sports, politics, technology etc. The features are used from the available dataset by the used classifier.

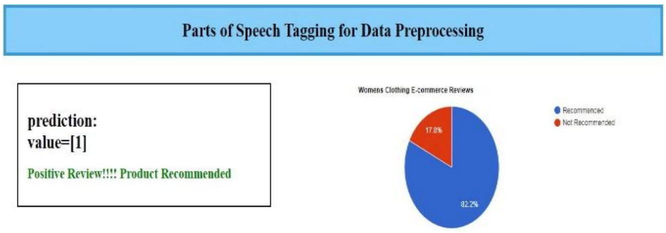
8. RESULTS

The proposed system provides the functionality of Text classification and POS tagging. The user input the text message to know the parts of speech, the number of tokens, elapsed time, polarity and subjectivity. In the case of reviews, the reviews are categorized into optimistic and pessimistic reviews through which the product is recommended or not to the next customers.



SL NO	Classification Algorithms	Accuracy	Confusion Matrix
1.	Decision Tree	96.3212%	[1915 47] [35 232]
2.	Multinomial NB	93.591%	[889 64] [155 7]
3.	Random Forest	97.219%	[953 0] [88 74]

Fig 8.1 Comparison of Algorithm Accuracy



Parts of Speech Tagging for Data Preprocessing

prediction:
value=[1]
Positive Review!!!! Product Recommended

Figure 8.2:- Prediction Mode

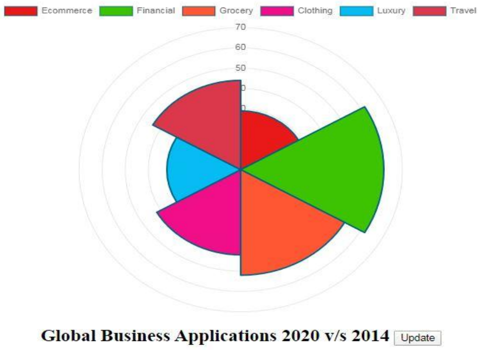


Fig 8.3:- Global Business Applications

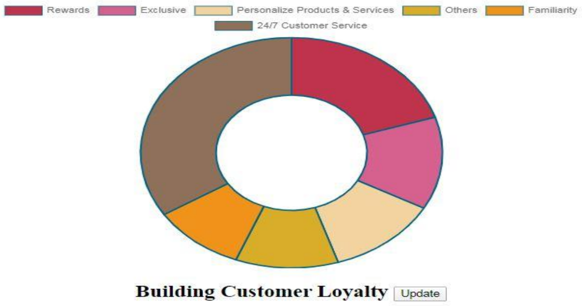


Fig 8.4:- Building Customer Loyalty

CONCLUSION

NLP takes a very important role in new machine human interface. When we look at some of the products based on the technologies of NLP we can see that they are very much advanced but very useful as well. This makes it very difficult and analyze. There are many languages spoken around the globe, thus it is difficult in building a model to predict accurate results. This problem gets more complicated when we think of different people speaking the same language but in different styles. Initially, domain independent model has been built and adapted directly with translation lexicons. This approach has boosted the probabilities of content words. This approach has further enhanced the probabilities of content words in the above proposed system. Significant improvements in perplexity have been observed in topic specific and domain independent models.

REFERENCES

[1] Parts-of-speech tagging based on dictionary & statistical machine learning 2016, 35th Chinese control

conference by Zhonglin Ye, Zhen Jia, Junfu Huang, Hongfeng Yin.

[2] POS Tagger Tokenizer by Saeed Rahmani, Mostafa Fakhra Ahmed published in 2015, 2nd International Conference on knowledge based Engineering.

ACKNOWLEDGEMENT

The research paper was completely supported by Don Bosco Institute of Technology. I am thankful to my guide Assistant Professor Mrs. Chaithra A S who provided extreme support that helped me in this research work.

BIOGRAPHIES

Umme Athiya
U G Student
Dept. of ISE
Don Bosco Institute of Technology
Bangalore, Karnataka

Mrs. Chaithra A S
Assistant Professor
Dept. of ISE
Don Bosco Institute of Technology
Bangalore, Karnataka