**IST 664 - Final Project Report**

**Group 8-Moring Batch**

**Detection of External References**

**Submitted by,**

**Rima Shah, Dhruva Pradhan, Siddesh Rane, Shreyas Sadagopan**

**Contents**

**Module 1: Introduction and Problem Statement**

**Module 2: Data Assimilation and Understanding**

**Module 3: Data Pre-Processing**

**Module 4: Exploratory Data Analysis**

**Module 5: Rule - Based Extraction: Creating a new Y Variable**

**Module 6: Feature Engineering**

**Module 7: Predictive Models**

**Module 8: Model Comparison - The original Y vs The rule based Y**

**Module 9: Inference & Conclusion**

**Module 1: Introduction and Problem Statement**

Reddit is a website where things like web content rating, social news aggregation and discussion takes place. The users can post content like links, images, etc. and then vote up or down for the posts put up by other users. Users can also comment on the posts posted by other users. We intend to study these submissions and comments to analyze the external sources present in these comments.

A model to predict whether a given text contains would be useful in many situations such as finding references in journals or papers, cross-referencing, etc. where we validate the presence of external sources to make sure the data is original. We can in turn use this to figure out truthfulness or validity of a text by calculating the number of external references cited. We can detect fake news through this algorithm too (If there are no supporting evidence for a given text).

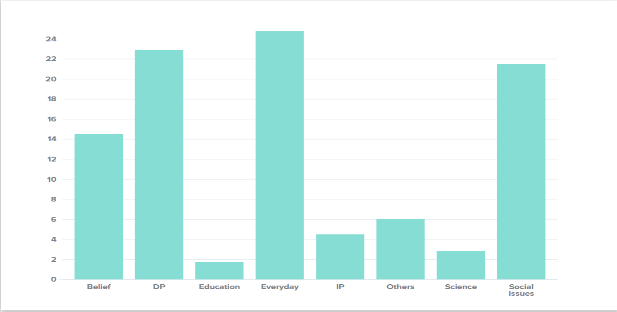
**Major Objectives**:

1. In this project, we will explore computational techniques to automatically identify if a given Reddit submission(comments) contains external sources or not. We will use the given sample dataset of Reddit submissions, to analyze the texts and build a supervised learning classifier
2. We conceive this as a classification problem where we would be classifying our text as containing an external source (1) or not containing an external source (0). We would be devising multiple feature sets- Unigram /bigram and a set of extra features. We would be exploring multiple classification algorithms and coming up with the best performing model to score our data set.
3. To create a Rule Based external references extractor. We would be using it as another Y variable to build another set of predictive models.
4. To compare the performances of all the models built and come up with a single best combination of the feature set, Y variable and the algorithm combination.

**Module 2: Data Assimilation and Understanding**

1. Dataset contains a list of 330 reddit submissions with entries covering varied topics of including education, social issues, science, politics etc.
2. Each row represents a single submission. Each submission is a lengthy paragraph and of varied lengths.
3. ‘Target Class’: In this data set there is a column with 1/0 values that tells if the comment in a record contains an external source or not. This value is the manually classified Y variable.
4. We then use the rule-based extraction algorithm to validate the manually classified Y variable and would be creating another Y variable. We would be using Naïve Bayes and Sklearn classifiers. We would be building 2 models, 1 with the initial Y variable and one rule base Y variable. The best performing model be the one with the highest accuracy and a good enough precision and recall values.

**Distribution of the initial Y variable Distribution of topics:**

**Module 3: Data Preprocessing**

1. Converted all the text to lower case using the lower() function

2. Removed punctuation marks from our texts

3. Removed words that contain numbers

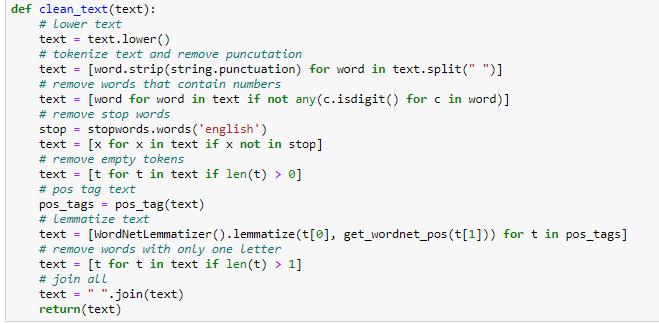
4. Removed stop words

5. Removed empty tokens

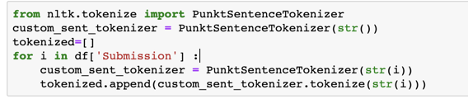
6. Lemmatized text with the WordNet() tagger

7. Removed words with only one letter

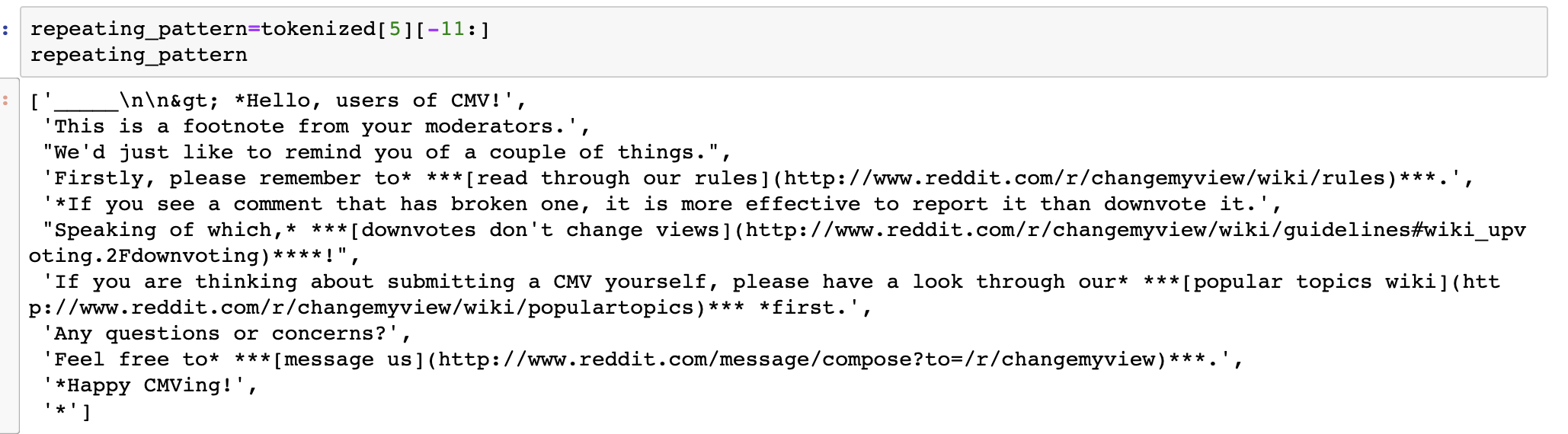
The below function takes care of the above steps:



8 Sentence tokenization: The preprocessing that we carried out first was tokening the comment into sentences. For tokenization we used PunktSentenceTokenizer’. This is an unsupervised tokenizer which uses abbreviation words, collocations, and words that start sentences. In most of the submissions, the last 11 sentences are junk so in the second step we removed those sentences.

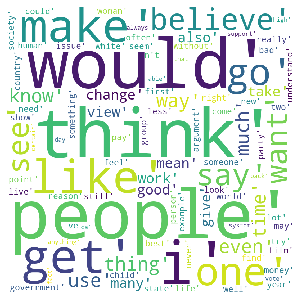


9. Removing text footers: We observed in the data that one footer line was attached to most of the submissions. We have removed this to make the texts consistent.

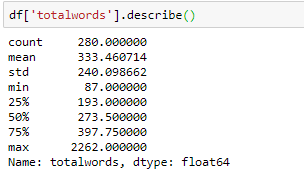


**Module 4: Exploratory Data Analysis**

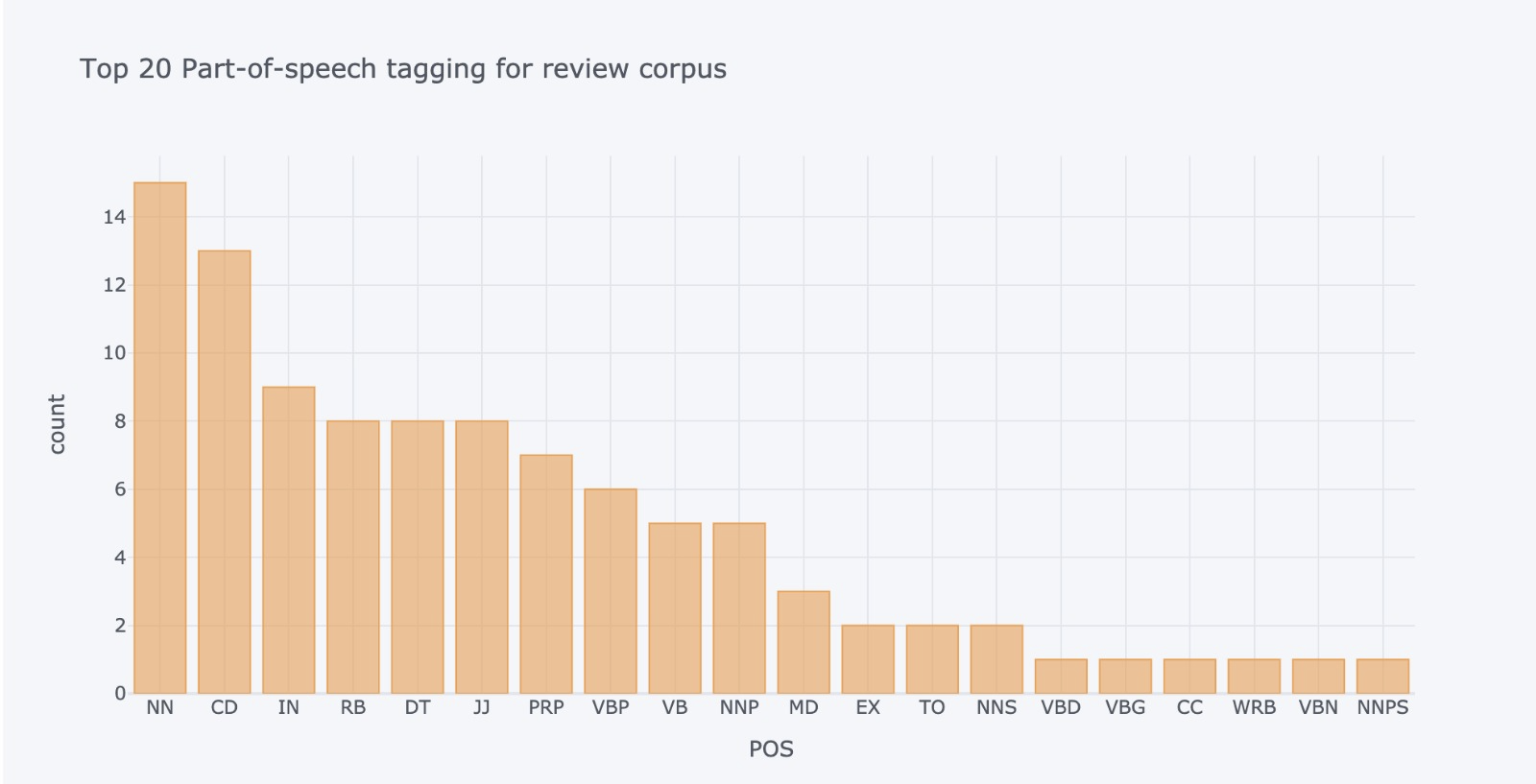
1. Visualized the top 100 most frequent words.



1. Distribution of the length of each of the submissions in the corpus: The average was around 334 words.



1. Top 20 parts of Speeches in the corpus



**Module 5: Rule - Based Extraction: Creating a new Y Variable**

The dataset in use currently has a Y variable which indicates whether a particular submission contains an External Reference (1/0). 1 indicates presence of an external Y variable. In order to validate the already present Y variable, we intend to create a new Y variable and develop models with each of these Y variables and analyze their performance.

1. **To develop a new Y variable, we have incorporated a Rule Based approach for the creation of the new Y variable. The following rules were used:**
2. **Contains an URL of an external website:**

Used a regex algorithm to filter out submissions with a URL. The following command checks for patterns in an URL, i.e., https, the back and forward slashes and .coms other than reddit.



This would in turn return a 1/0 output flagging submissions with an external URL.

|  |  |  |
| --- | --- | --- |
|  | **Total Number** | **With URL** |
| **Number of Submissions** | 330 | 46 |

1. **Contains references inside square brackets:**

We filter for submissions that contain a square bracket and we count the number of words and analyze the type of words inside the square brackets.

If the following patterns are found inside the square brackets, we flag that submission as 1:

1. Number of words inside brackets>5
2. Contains numbers like [1], [2] and [3] etc.
3. Contains the pattern (topic: topic name)
4. Contains [article]

|  |  |  |
| --- | --- | --- |
|  | **Total Number** | **With references inside square brackets** |
| **Number of Submissions** | 330 | 55 |

1. **Contains ‘said’ like words near pro-nouns/ the source:**

We filter using a set of regexes that return true if the word ‘said’ like word appear within 10 words of a pronoun, capitalized (proper) noun and possible speakers *( he/she/ journals/ TV Channels/ they/ )* etc.

1. ‘Said’ like words include: Say, report, reported, quote, tell, told, said, added, exclaimed, exclaims, and says.
2. We tried to stem the sentences and apply this rule. But we ended up capturing words that are not necessarily quotations.
3. We also excluded patterns like ‘being said’, ‘having said’ and ‘I Said’ which are not external references
4. If the submission gets captured by this following command, it would be flagged as 1.

|  |  |  |
| --- | --- | --- |
|  | **Total Number** | **Contains ‘said’ like words near pro-nouns** |
| **Number of Submissions** | 330 | 166 |

1. **Contains more than 20 words inside quotes:**

Average number of words inside double quotes: 12 words

Average number of words inside double quotes which are external quotations: 20 words

We filter for submissions that contain double quotes and we count the number of words and analyze the type of words inside the square brackets.

If the following patterns are found inside the square brackets, we flag that submission as 1:

1. Number of words inside brackets > 20
2. Appears after ‘said’ like words.
3. Appears after a semicolon.

|  |  |  |
| --- | --- | --- |
|  | **Total Number** | **Contains more than 20 words inside quotes:** |
| **Number of Submissions** | 330 | 67 |

1. **Overlap between all the rules: Unique submissions with external references**

We found the unique submissions which were filtered using the above mentioned 4 rules. We deleted the duplicates and got our required Y variable. These are the set of submissions with external references

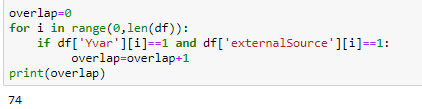
|  |  |  |
| --- | --- | --- |
|  | **Total Number** | **Submissions with external references (Rule Based)** |
| **Number of Submissions** | 330 | 124 |

1. **Rules/Techniques we tried out and which we intend to do in the future:**
2. Creating POS tag-based extractor which would identify patterns in external reference:
   1. Parts of speeches in external references (Count/Average/Mode)
   2. Parts of speeches of words that trail/follow an external reference
3. Create a rule based chunker model to identify grammars and nodes in a reference. (We tried to do this, but were not successful)
   1. If we can try to map quotes with their speakers/ publishers, we can conclude that the submission contains an external source.
4. **Comparing the created Y variable, with our parent Y variable**.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Total Number** | **Submissions with external references (Initial Y Variable)** | **Submissions with external references (Rule Based)** |
| **Number of Submissions** | 330 | 91 | 124 |

**Overlap Between The Original Y vs The Rule Based Y: 74 submissions**

**Correlation Between The Original Y and The Rule Based Y: 68%**



**Module 6: Sampling for Our Predictive Models**

1. We need to classify a given submission as a ‘1’ or ‘0’ using our Y variables (The initial and the rule based)
2. Our total number of 1 VS 0 submissions in both the Y variables are unbalanced: We need to make the dataset balanced for our predictive models

**A. Set 1: The Initial Y Variable sampling**

* 1. Before Sampling: Number of 1s = 91, Number of 0s = 229



* 1. After sampling: Number of 1s= 91, Number of 0s=91



**B**. **Set 2: The Rule-Based Y variable**

1. Before Sampling: Number of 1s = 124, Number of 0s = 206

****

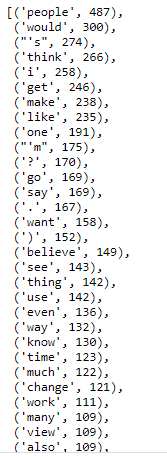
1. After Sampling: Number of 1s = 124, Number of 0s = 124



**Module 6: Feature Engineering:**

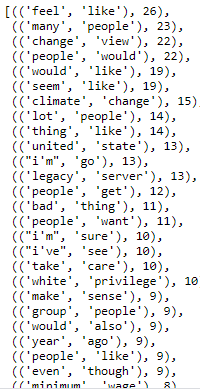
Created 4 different features for each of the under sampled data sets- 1 and 2 from the above module

1. **Bag of Words:** 
   1. Created a Bag of Words feature set for the above under sampled data set.
   2. Filtered for the top 50 frequent words in the corpus: Have not cleaned the alphanumeric as their count has an effect on our output



* 1. Mapped the above words throughout the corpus and created a feature set each for each of the data sets
     1. This has the true of false values for each of the above words and the Y variable for each of the submissions in the corpus

1. **Bigram Features**
   1. Created a Bigram feature set for the above under sampled data sets -1 and 2.
   2. Filtered for the top 50 frequent bigram combinations in the corpus: Have not cleaned the alphanumeric as their count influences our output



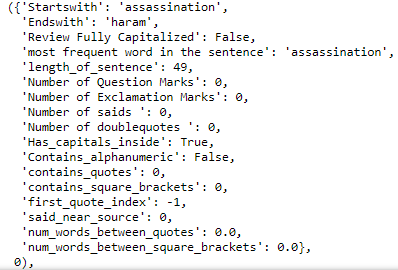
* 1. Mapped the above combinations throughout the corpus and created a feature set each for each of the data sets
     1. This has the true of false values for each of the above words and the Y variable for each of the submissions in the corpus

1. **Extra Features:**

**Created feature sets with the following features:**

1. Starting word of the Submission
2. Ending word of the Submission
3. Length of the Submission
4. Is the submission in Full Capitals
5. Does the submission contain capitalized words in between?
6. Number of Question Marks
7. Number of Exclamation marks
8. Does it contain an Alphanumeric
9. Contain Quotes (1/0)
10. Contains Square Brackets (1/0)
11. First Quote Index
12. Does it contain ‘said’ like features near quotes
13. Number of Words between Quotes
14. Number of Words Between Sq.Brackets
15. Number of Saids
16. Number of double quotes

**The above variables were mapped with the 2 data frames and the resulting feature set is below:**



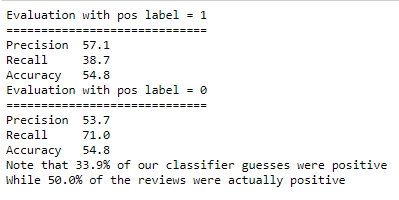
**Module 7: Predictive Models**

To predict whether a given submission contains external sources, we have created 2 sets of predictive models, one with the initial Y variable and one with the rule based derived Y variable. We have built the predive models using the 2 of the NLTK classifiers: Naïve Bayes and Sklearn Classifier.

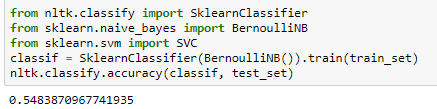
For building the classification model, we followed incremental techniques and built a base model in each of the cases and tuned them using ‘Hyperparameter tuning’ to arrive at the best parameter combinations. We build models using each of the feature sets and choose the best model for our case.

**SET 1: Models using the Initial Y variable**

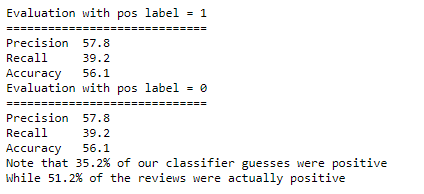
* 1. **Feature set 1: Bag of Words**
     1. **Naïve Bayes Classifier: Got an Accuracy of 55%**



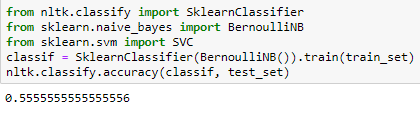
* + 1. **Sklearn Classifier: Got an Accuracy of 54%**



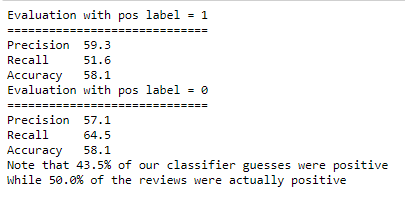
* 1. **Feature Set 2: Big Gram Model**
     1. **Naïve Bayes Classifier: Got an Accuracy of 56.1%**



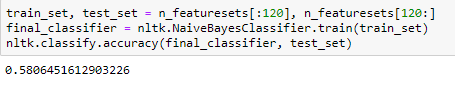
* + 1. **Sklearn Classifier: Got an Accuracy of 55.6%**



* 1. **Bag of Words+ Bigrams and The Extra features**
     1. **Naïve Bayes Classifier: Got an Accuracy of 58.1%**

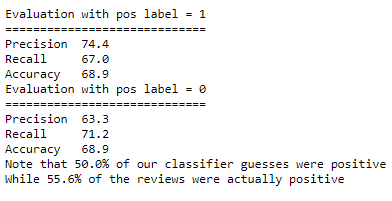


* + 1. **Sklearn Classifier: Got an Accuracy of 58.06%**

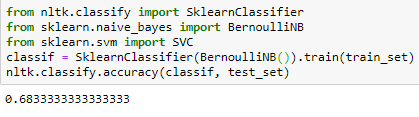


**SET 2: Models using the Rule based Y variable**

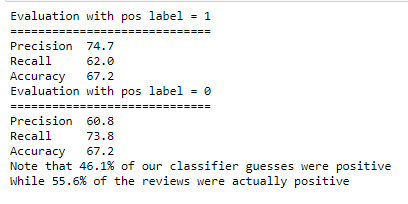
1. **Feature set 1: Bag of Words**
2. **Naïve Bayes Classifier: Got an Accuracy of 68.9%**



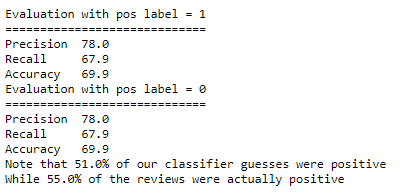
1. **Sklearn Classifier: Got an Accuracy of 68.3%**



1. **Feature Set 2: Big Gram Model**
2. **Naïve Bayes Classifier: Got an Accuracy of 67.2%**



1. **Sklearn Classifier: Got an Accuracy of 67.5%**
2. **Bag of Words+ Bigrams and The Extra features**
3. **Naïve Bayes Classifier: Got an Accuracy of 69.9%**



1. **Sklearn Classifier: Got an Accuracy of 69.5%**

**Module 8: Model Comparison - The original Y vs The rule based Y**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Models with the Initial Y variable** | | **Models with the Initial Y variable** | |
| **Feature set** | **Sklearn Accuracy(%)** | **Naïve Bayes Accuracy(%)** | **Sklearn Accuracy(%)** | **Naïve Bayes Accuracy(%)** |
| Bag of words | **54** | **55** | **68.3** | **68.9** |
| Bigram Features | **55.6** | **56.1** | **67.5** | **67.2** |
| **Extra Features**+ Bag of Words+ The Bigram features | **58.06** | **58.1** | **69.5** | **69.9** |

From the numbers above, we can conclude that the Rule base Y variable we created provides a greater accuracy when use in predictive models. Our rule base Y variable has a greater number of 1’s and hence understandably has more recall and precision values too. From the above table of accuracies, we can conclude that the best performing model is with 69.9% accuracy the Naïve Bayes Model. We get the best accuracy while using all our feature sets.

**Y variable to Use**: Rule Based Y variable

**Feature sets to use**: Extra Features+ Bag of Words+ Bigram features

**Classifier to use:** Naïve Bayes Algorithm

**This model has a great recall and precision values too.**

**Module 9: Inference & Conclusion**

To conclude and to answer our major objectives:

1. We were able to explore and identify computational techniques to automatically determine if a given Reddit submission(comments) contains external sources or not. We have built multiple predictive models to determine this.
2. We have built multiple feature sets including the unigram bag of words, bigram features and some extra features for our use case.
3. We have created rules to extract references from a given text. Using this rule base extractor, we have created predictive models to automatically predict if the given text contains an external reference or not.
4. We have analyzed all both the model sets (using the initial Y and the rule-based Y) and have drawn some comparative studies between them. It was seen that the Rule based Y model performs the best when used with all the 3 feature sets in tandem
5. We have thus built an algorithm to automatically classify a given text based on whether it contains an external reference.

**Module 10 : References :**

Hitkul, Karmanya Aggarwal, Neha Yadav, and Maheshwar Dwivedy. "A Comparative Study of Machine Learning Algorithms for Prior Prediction of UFC Fights." *Harmony Search and Nature Inspired Optimization Algorithms Advances in Intelligent Systems and Computing* (2018): 67-76.

Johnson, Jeremiah Douglas. *Predicting outcomes of mixed martial arts fights with novel fight variables*. Diss. University of Georgia, 2012.

Kuchen, Robert. "Model Based Prediction of Starcraft II Match Outcomes."

El-Fouly, T. H. M., E. F. El-Saadany, and M. M. A. Salama. "Improved grey predictor rolling models for wind power prediction." *IET Generation, Transmission & Distribution* 1.6 (2007): 928-937.

Soman, Saurabh S., et al. "A review of wind power and wind speed forecasting methods with different time horizons." *North American Power Symposium 2010*. IEEE, 2010.