|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Number of Doc** | **Threshold** | **Accuracy** |
| Voting Full Document | 20 | Threshold : 1.6000003 | Accuracy : 0.65 |
|  | *40* | *Threshold : 1.4000002* | *Accuracy : 0.75* |
| 60 | Threshold : 1.3000002 | Accuracy : 0.68333334 |
| 100 | Threshold : 1.6000003 | Accuracy : 0.66 |
| Adjective Count Full Document | 20 | Threshold : 0.5 | Accuracy : 0.75 |
|  | *40* | *Threshold : 0.3* | *Accuracy : 0.8* |
|  | 60 | Threshold : 0.5 | Accuracy : 0.71666664 |
|  | 100 | Threshold : 0.5 | Accuracy : 0.75510204 |
| CueBased Full Document | *20* |  | *Accuracy : 0.5* |
|  | *40* |  | *Accuracy : 0.5* |
|  | *60* |  | *Accuracy : 0.5* |
|  | 100 |  | Accuracy : 0.48979592 |
| OtherPOS Full Document | 20 | Threshold : 0.7 | Accuracy : 0.7 |
|  | *40* | *Threshold : 0.7* | *Accuracy : 0.725* |
|  | 60 | Threshold : 0.9000001 | Accuracy : 0.68333334 |
|  | 100 | Threshold : 0.6 | Accuracy : 0.6020408 |
| Voting FirstSentence | *20* | *Threshold : 0.2* | *Accuracy : 0.6* |
|  | 40 | Threshold : 0.2 | Accuracy : 0.55 |
|  | 60 | Threshold : 0.2 | Accuracy : 0.51666665 |
|  | 100 | Threshold : 0.2 | Accuracy : 0.5408163 |
| Adjective Count FirstSentence | *20* | *Threshold : 0.1* | *Accuracy : 0.6* |
|  | 40 | Threshold : 0.1 | Accuracy : 0.525 |
|  | *60* | *Threshold : 0.1* | *Accuracy : 0.6* |
|  | 100 | Threshold : 0.3 | Accuracy : 0.52040815 |
| OtherPOS FirstSentence | 20 | Threshold : 0.5 | Accuracy : 0.5 |
|  | 40 | Threshold : 0.5 | Accuracy : 0.5 |
|  | *60* | *Threshold : 0.1* | *Accuracy : 0.533* |
|  | 100 | Threshold : 0.2 | Accuracy : 0.52040815 |
| Voting LastSentence | *20* | *Threshold : 0.1* | *Accuracy : 0.7* |
|  | *40* | *Threshold : 0.1* | *Accuracy : 0.7* |
|  | 60 | Threshold : 0.1 | Accuracy : 0.68333334 |
|  | 100 | Threshold : 0.1 | Accuracy : 0.67346936 |
| Adjective Count LastSentence | *20* | *Threshold : 0.1* | *Accuracy : 0.6* |
|  | 40 | Threshold : 0.1 | Accuracy : 0.575 |
|  | 60 | Threshold : 0.1 | Accuracy : 0.56666666 |
|  | 100 | Threshold : 0.1 | Accuracy : 0.56122446 |
| OtherPOS LastSentence | 20 | Threshold : 0.2 | Accuracy : 0.55 |
|  | *40* | *Threshold : 0.1* | *Accuracy : 0.625* |
|  | 60 | Threshold : 0.1 | Accuracy : 0.6 |
|  | 100 | Threshold : 0.1 | Accuracy : 0.5816327 |
| Adjective Count Significant | *20* | *Threshold : 0.2* | *Accuracy : 0.85* |
|  | 40 | Threshold : 0.2 | Accuracy : 0.75 |
|  | 60 | Threshold : 0.1 | Accuracy : 0.76666665 |
|  | 100 | Threshold : 0.2 | Accuracy : 0.81632656 |
| OtherPOS Significant | *20* | *Threshold : 0.6* | *Accuracy : 0.5* |
|  | *40* | *Threshold : 0.6* | *Accuracy : 0.5* |
|  | *60* | *Threshold : 0.6* | *Accuracy : 0.5* |
|  | 100 | Threshold : 0.6 | Accuracy : 0.5102041 |
| Adjective Count All(FLS) | *20* | *Threshold : 0.2* | *Accuracy : 0.85* |
|  | 40 | Threshold : 0.2 | Accuracy : 0.725 |
|  | 60 | Threshold : 0.2 | Accuracy : 0.76666665 |
|  | 100 | Threshold : 0.2 | Accuracy : 0.82653064 |
| OtherPOS All(FLS) | 20 | Threshold : 0.5 | Accuracy : 0.5 |
|  | *40* | *Threshold : 0.1* | *Accuracy : 0.55* |
|  | *60* | *Threshold : 0.1* | *Accuracy : 0.55* |
|  | 100 | Threshold : 0.1 | Accuracy : 0.5408163 |

**7. CONCLUSION:**

The main aim of the project was to classify given text as opinion or factual using a threshold based approach. To achieve this goal the path followed can be compactly concluded as below:

* Collection of data: Collection of several sets of input text files. The sets consisted of both opinionated and factual text documents.
* Implementation: Identifying and implementing different methods of threshold based classification with different input types. The following list shows the classification methods experimented along with the input types they support:
  + Voting based – Count of polar words in sentence(s). Supports input types – Full document, First Sentence, last Sentence.
  + Adjective Count Based – Count of adjectives in POS tagged sentence(s). Supports input types - Full document, First Sentence, last Sentence, Significant Sentence, First-last-significant sentence.
  + Cue Based – find opinion oriented cues in the document. Supports input types – Full document
  + Other POS – find opinion oriented POS patterns in the document. Supports input types - Full document, First Sentence, last Sentence, Significant Sentence, First-last-significant sentence.
* Experimentation: Execution of all the above mentioned methods with inputs consisting of varying number of text documents. The results were recorded with a range of thresholds and their accuracies for each method experimented.
* Visualization of results as graphs.
* Inference: Study of the results and determination of the most accurate method to classify any input text. Also observing the behavior of the accuracy of classification against number of input files as a factor.
* Front End: Creation of front end for user interaction with three modules – to view experiment results, to classify user entered text as opinion or fact and to view the accuracy of classification against number of input files.
* Documentation: Detailed report of the project consisting of explanation of methods and phases followed during building of the project.

The project has been to determine an optimal classification technique with a threshold giving satisfying accuracy. This can be applied in the process of extracting opinion oriented text documents from large collection data documents, which can serve as input to other opinion mining applications like opinion-search engines etc.,

Inferences:

Voting Full Document : Threshold ranges around 1.5 give high accuracies. Indicating that documents that contain sentences with average 1.5 or greater as number of polar words per sentence are more probable to be opinion oriented.

Voting First Sentence: Threshold ranges around 0.2 give high accuracies. Indicating that documents with their first sentence containing 20 percent or more words to be polar are more probable to be opinion oriented.

Voting Last Sentence: Threshold ranges around 0.1 give high accuracies. Indicating that documents with their first sentence containing an average 10 percent or greater words to be polar are more probable to be opinion oriented.

Adjective Count Full Document: Threshold ranges around 0.5 give high accuracies. Indicating that documents that contain sentences with average 0.5 and greater as number of adjectives per sentence are more probable to be opinion oriented.

Adjective First Sentence: Threshold ranges around 0.1 give high accuracies. Indicating that documents with their first sentence containing an average 10 percent and higher number of words to be adjectives are more probable to be opinion oriented.

Adjective Last Sentence: Threshold ranges around 0.1 give high accuracies. Indicating that documents with their last sentence containing an average 10 percent and higher number of words to be adjectives are more probable to be opinion oriented.

Adjective Significant Sentence: Threshold ranges around 0.2 give high accuracies. Indicating that documents with their significant sentence containing an average 20 percent or higher number of words to be adjectives are more probable to be opinion oriented.

Adjective First-Last-Significant: Threshold ranges around 0.2 give high accuracies. It considers the contribution of the first last and significant sentences in the text. This method gives a high accuracy since it’s scope of consideration is more and very relevant.

Cue Based Full Document: Gives an accuracy of 50%. Indicates that presence of cues give a 50% probability of the text being opinionated.

Other POS Full Document: Threshold ranges around 1 give high accuracies. Indicating that documents that contain sentences with average 1 and greater as number of matching POS patterns per sentence are more probable to be opinion oriented.

Other POS First Sentence: Threshold ranges around 0.1 to 0.2 give high accuracies. This matches patterns only in the first sentence.

Other POS Last Sentence: Threshold ranges around 0.1 give high accuracies. This matches patterns only in the last sentence.

Other POS Significant Sentence: Threshold ranges around 0.5 give high accuracies. This matches patterns only in the significant sentence.

OtherPOS First-Last-Significant: Threshold ranges around 0.3 give high accuracies. This matches patterns in the first, last and significant sentences.