Auto Text Summarization with Categorization and Sentiment Analysis

Ashmita Shetty
Student, M.E. (Computer Engineering)
Thadomal Shahani Engineering College
Mumbai, India

Ruhi Bajaj
Assistant Professor, Computer Department
Thadomal Shahani Engineering College
Mumbai, India

ABSTRACT
In today's world the volume of information is dramatically increasing, and the value of that information is growing fast. Modern organizations deals with terabytes of text, such as email, that often plays a significant role in their day-to-day operations. Even small and medium-sized organizations are dealing with growing volumes of text that require rapid access and meaningful analysis. Identifying useful information from these data is quite difficult and requires some mechanism. One possible means is to use text categorization and summarization. Text categorization is automatically arranging a set of documents into predefined categories and Summarization is a giving a condensed and precise depiction of input data such that the output includes the most significant concepts of the source. Sentiment analysis i.e. opinion mining states the use of NLP, text analysis and to identify and extract biased information in source materials.

General Terms
Text Mining, Fuzzy Systems, Natural Language Processing, Sentiment Analysis.

Keywords
Categorization, Feature matrix, Fuzzy Logic, Sentiment analysis.

1. INTRODUCTION
With massive growth of information on WWW, Conventional Information Retrieval techniques have become unproductive for finding relevant information effectively. Given a keyword-based search on the internet, it returns thousands of documents overwhelming the user. It becomes very challenging and time consuming task to find the significant documents. This in turn requires the user to analyze the searched results one by one until satisfied information is acquired, which is time-consuming and inefficient. It is therefore essential to develop tools to efficiently assist users in identifying desired documents.

Text Categorization and Summarization is done on the input documents. After obtaining the summary of the document, sentiment analysis is done on it to identify whether the result of the summary is positive or negative.

Automatic text categorization has always been a vital application. A text categorization is used in ordering documents to support information retrieval tasks. The text classification task can be defined as assigning category to new documents based on the knowledge gained.

The role of summarization is to present the most important information from the text in the shorter version without changing meaning of the original text. Summarization can be classified in two types Extraction and Abstraction. Extraction of the document is the selection of sentence that has highest score among other document. Wherein abstraction involve use of linguistic method and extract the sentence together to constitute something new, that is not present in the source, and substitute them in the summary with new concept.

Sentiment analysis helps to evaluate ideas, feelings, attitude and behavior, which is used to make decisions. A simple task in sentiment analysis is categorizing the polarity of a given text in the document, whether the expressed sentiment in a document is positive or negative. It is not only helps the people, but also helps the company to evaluate sentiment or opinions and behavior of the customer, who is using their products and can get an opinion about its product which helps organizations during the decision making process.

The system describes mainly three parts:

1. Categorization
2. Fuzzy System
3. Sentiment Analysis.

The categorization is based on term weight frequency and accordingly the documents are categorized into predefined categories. Pre-Processing including three major procedures namely:

- Sentence Segmentation
- Removing Stop Words
- Removing stemming Words

The fuzzy system is provided with extraction of features of document based on:

- Title feature
- Sentence length
- Term weight
- Sentence position
- Proper noun

Once the feature matrix is obtained, each sentence of the document is represented by the sentence score. Sentence having the highest score is extracted to form the summary. And finally Sentiment is generated for the summary.

2. LITERATURE SURVEY
A document summarization framework is proposed via deep learning using fuzzy logic. [1] The framework consists of concepts extraction, summary generation and reconstruction validation. A query-oriented extraction technique focus on information that are distributed in multiple documents to hidden components layer by layer. In text summarization the document is preprocessed using various preprocessing techniques and then it is transformed into feature matrix. This feature matrix, each row will work as a input to Restricted

Boltzmann Machine. [1] The defuzzifier output is given as input to the deep learning algorithm where it uses bias values which are randomly selected.

A Sentiment Summarization system takes as input a set of documents that contain opinions about some entity of interest. [2] Subsequently, it processes all the given documents and generates a summary of all the input documents. This summary should represent the average view of all the documents and important aspects of the target of SA addressed in those documents. Sentiment summarization consists of four main steps. Firstly, finding the target item's features (e.g. battery in the context of smart phones) that most people have commented on in the corpus of tweets. Next, grouping the documents containing the same item feature in the same cluster. Third, classifying those documents using a polarity detection system. Finally, output a textual, as well as, a graphical summary, based on the classified documents.

Text categorization which is the task to classify a document under predefined category. [3] Here, the document is summarized and the two approaches for text categorization is based on the features extracted using KNN algorithm. [3] Preprocessing is to convert the semi-structured text into Vector Space Model (VSM). Automatic summarizations, weight of verb and noun, distance between words, sentence importance, etc are calculated. Based on these the features are selected. Classifier then calculates based on these features.

3. METHOD

The implementation plan consists of three parts, categorizing phase where the documents will be categorized and then the summarization phase where based on the categories the summary is generated and finally the sentiment analysis is carried out.

3.1 Categorization

The documents that were loaded are selected to be classified into various predefined categories. It is done using Term Frequency which measures how often a term occurs in a document. [4]

3.2 Preprocessing

The following sections present the pre-processing of the input dataset, which will give all features: [5]

3.2.1 Sentence Segmentation

Sentence segmentation is performed by identifying the delimiter commonly denoted by "." called as full stop. It is used to separate the sentences in the document.

3.2.2 Stopword Removal

Stop words removal is the process of removing words which do not convey any meaning during the classification process.

3.2.3 Stemming

Stemming is a method to reduce a word to its stem or root form. Porter's stemming algorithm is used for this purpose.

3.3 Feature Matrix

The whole documents under consideration are subjected for the feature extraction and a set of features are extracted accordingly. Based on these features the feature values will be assigned. The feature matrix is constructed according to the sentences extracted from the multiple documents. The feature sets are:

3.3.1 Title Feature

It is the ratio of the number of words in the sentence that occur in title to the total number of words in the title helps to calculate the score of a sentence for this feature. [6]

$$Title\ feature\ (f1) = \frac{No\ of\ words\ \ in\ sente\ nce}{No\ of\ words\ \ in\ Title} \quad (1)$$

3.3.2 Sentence Length

Length of the sentence is the proportion of the number of words arising in the sentence over the number of words arising in the longest sentence of the document. [7]

3.3.3 Term Weight

The Total Term Weight is calculated by Term Frequency and IDF for a document. It is obtained by dividing the total number of documents by the number of documents holds the term.

$$f_3 = TF \times IDF(t,d,D) = TF(t,d) \times IDF(t,D) \quad (2)$$

3.3.4 Proper Noun

The sentence that holds more proper nouns (name entity) is an essential and it is most probably included in the document summary. [7] It is a ratio of number of proper noun in the sentence to the length of the sentence.

$$f_5 = \frac{No\ of\ Proper\ \ noun\ \ in\ sentenc\ \ e}{Sentence\ \ \ Lengt\ h} \quad (3)$$

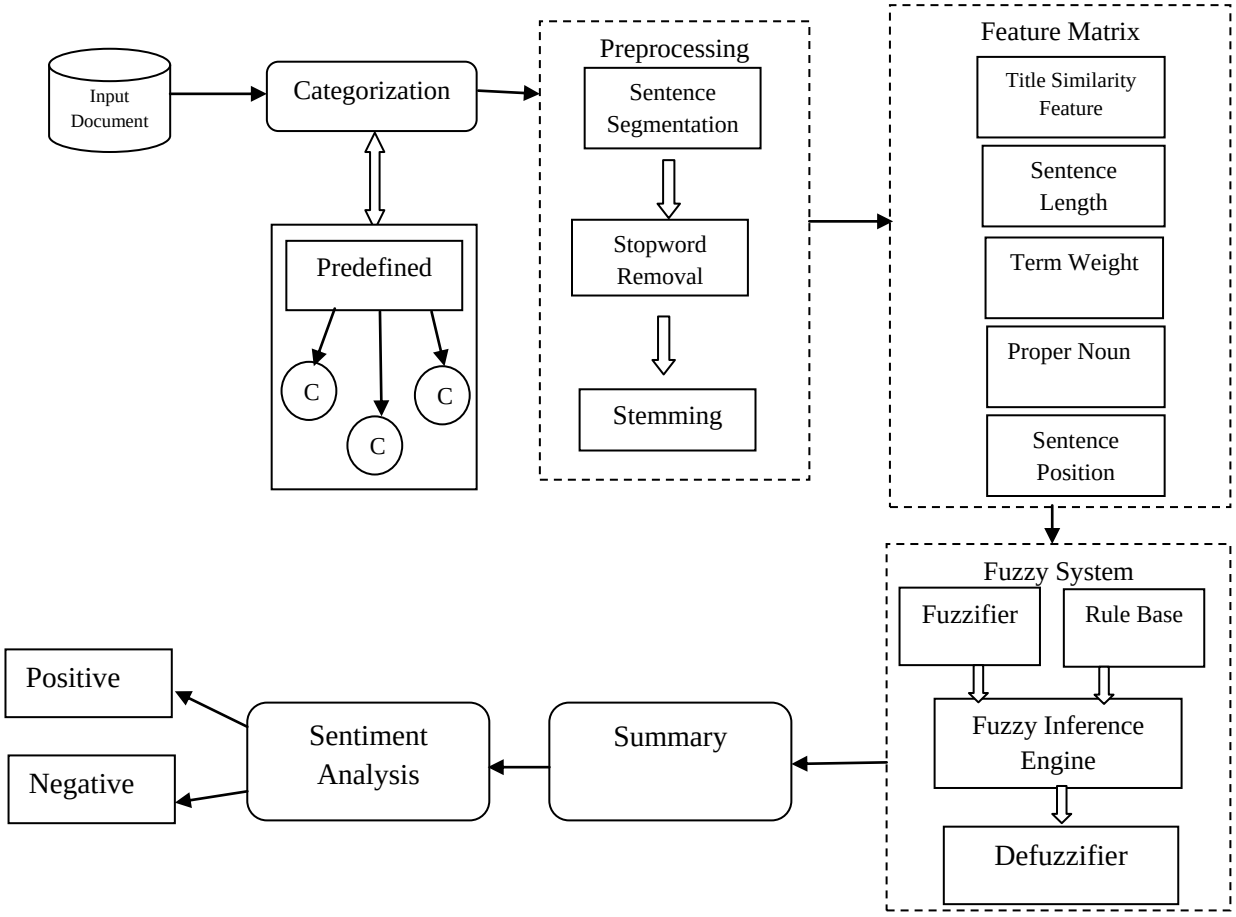

Fig 1: System Diagram

3.3.5 Sentence Position

If the sentence given is in the starting of the sentence or the last in the sentence of the paragraph then the feature value f2 is assigned as 1. Else if the sentence is in the middle of the paragraph then the feature value of f2is assigned as 0. [8]

3.4 Fuzzy Logic

The fuzzy logic system consists of four components: fuzzifier, inference engine, defuzzifier and the fuzzy knowledge base. [9] In order to implement text summarization based on fuzzy logic, first, the five features extracted in the previous section are used as input to the fuzzifier. The input data uses triangular membership function for each feature that is divided into fuzzy sets as low (L), medium (M) and High (H). Inference engine the most significant part is the definition of fuzzy IF-THEN rules. [9] The important sentences are extracted from these rules according to our features criteria.

For Example:

IF (SentenceLength is H) and (TermFreq is H) and (SentencePosition is H) and (NoProperNoun is M) THEN (Sentence is important)

The last step is the defuzzfication. The output membership functions are: Unimportant, Average, and Important which is used to convert the fuzzy results into a crisp output for the final score of each sentence. [10]

3.5 Sentiment Analysis

In order to determine the sentiment polarity of an adjective describing an opinion feature we make use of SentiWordNet which is a lexical resource for opinion mining. [11] SentiWordNet assigns normalized sentiment scores: positivity and negativity to each synset of WordNet. The adjective and adverb from the summary are analyzed for their positive and negative polarity and the result is generated. POS Tagging is done to identify the words in the summary. [11] To calculate the sentiment analysis of the summary, the following formula is implemented:

$$S(T)=\sum[S(AVG)+s(AG)] \quad (4)$$

Where,

S(AVG) denotes score of adverb

S(AG) denotes score of adjective

4. RESULTS AND DISCUSSION

The system is developed for categorization, summarization and sentiment analysis using dot net technology. Text categorization is done using Term Frequency and summarization is done using Fuzzy System and finally sentiment analysis is carried out using SentiWordNet. The system starts by categorizing the documents into predefined categories and the preprocessing technique is applied to extract features from the document, the features are assigned scores for the sentence in the document. Using these features the fuzzy system is feed to get the required summary.

Sentiment analysis is carried out on the document summary to identify if it is positive or negative.

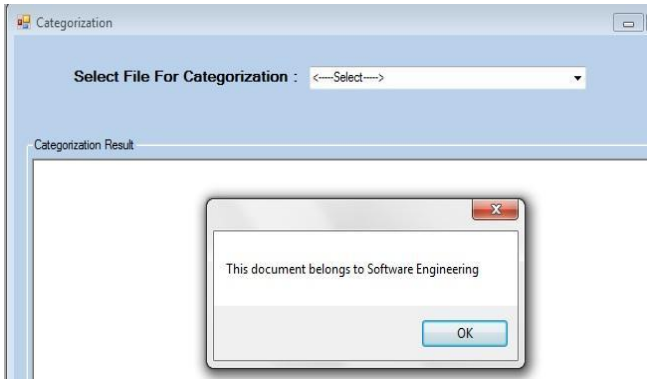

Fig 2: Sample Category for the document selected

Fig 3: Sample Summary generated

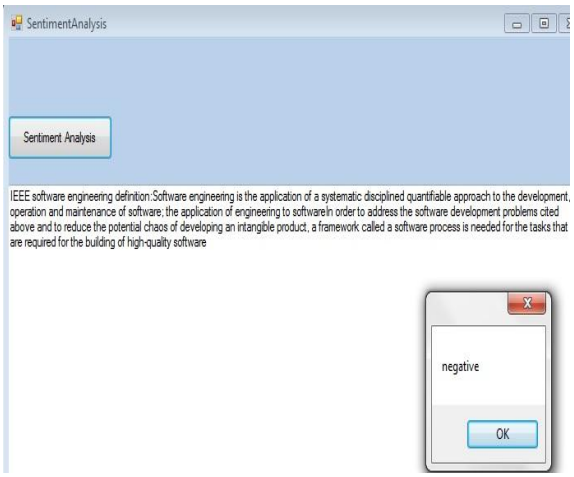

Fig 4: Sentiment analysis for the sample document summary

5. CONCLUSION

In this system, an automated text categorization, text summarization and sentiment analysis is implemented. Text categorization has currently four categories. The summarization is based on feature extraction that uses fuzzy logic. The system uses five features to get the summary. The system is tested for 20 documents of around 90 words each. The summary provides the complete information about the document. The quality of the summary can still be improved by increasing the number of features. Finally sentiment analysis uses SentiWordNet to get the values for the words and based on the score of the noun and adverb the sentiment is analyzed.

The system can be further improvised to include more languages and could also translate the summary into various other languages. The quality of the summary can also be enhanced by extracting more features during the preprocessing phase.

6. REFERENCES

[1] G. Padmapriya, Dr. K. Duraiswamy, "Association of Deep Learning Algorithm with Fuzzy Logic for Multidocument Text Summarization", *Journal of Theoretical and Applied Information Technology*, April 2014.

[2] Seyed-Ali Bahrainian, Andreas Dengel, "Sentiment Analysis and Summarization of Twitter Data", *2013 IEEE 16th International Conference on Computational Science and Engineering*, 2013.

[3] JIANG Xiao-Yu, FAN Xiao-Zhong, Wang Zhi-Fei,JiaKe-Liang, "Improving the Performance of Text Categorization using Automatic Summarization", *International Conference on Computer Modeling and Simulation*, 2009 IEEE.

[4] V. Gupta, G. S. Lehal, "A Survey of Text Mining Techniques and Applications", *Journal of Emerging Technologies in Web Intelligence*, Vol. 1, No. 1, August 2009.

[5] C. Ramasubramanian, R. Ramya, "Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2, Issue 12, December 2013.

[6] L. Suanmali, N. Salim, M.S. Binwahlan, "SRL-GSM: A Hybrid Approach based on Semantic Role Labeling and General Statistic Method for Text Summarization", *Research Article- Journal of Applied Science*, 2010.

[7] M. K. Dalal, M. A. Zaveri, "Semisupervised Learning Based Opinion Summarization and Classification for Online Product Reviews", *Hindawi Publishing Corporation Applied Computational Intelligence and Soft Computing*, Volume 2013.

[8] A. Kiani, M. R. Akbarzadeh, "Automatic Text Summarization Using: Hybrid Fuzzy GA-GP", *International Conference on Fuzzy Systems*, 2006 IEEE.

[9] R. D. Shinde, S. H. Routela, S. S. Jadhav, S. R. Sagare, "Enforcing Text Summarization using Fuzzy Logic", *International Journal of Computer Science and Information Techologies*, Volume 5(6), 2014 IJCSIT.

[10] L. Suanmali, N. Salim, Mohammed Salem Binwahlan, "Feature-Based Sentence Extraction Using Fuzzy Inference rules", *International Conference on Signal Processing Systems*, 2009 IEEE.

[11] S. Rana, "Sentiment Analysis for Hindi Text using Fuzzy Logic", *Indian Journal of Applied Research*, Volume 4, Issue 8, August 2014.