IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

NOVEL AUTOMATIC TEXT CLASSIFICATION TECHNIQUE USING NATURAL LANGUAGE PROCESSING

[1]K. S. Vanishree, [2]Manjunatha. H. R.
[1,2]Assistant Professor, Department of Computer Science, Government First Grade College, Shivamogga, Karnataka, India.
Email ID's: hrmanjunath.spr@gmail.com, vanishree.kss@gmail.com

ABSTRACT: In this informative age, many documents in different Indian Languages are available in digital forms. For easy retrieval of these digitized documents, these documents must be classified into a class according to its content. Text Classification is an area of Text Mining which helps to overcome this challenge. Text Classification is act of assigning classes to documents Text mining and natural language processing are fast growing areas of research, with numerous applications in business, science and creative industries. Automated text classification has been considered as a vital method to manage and process a vast amount of documents in digital forms that are widespread and continuously increasing. In general, text classification plays an important role in information extraction and summarization, text retrieval, and question answering. The demand of text classification is growing significantly in web searching, data mining, web ranking, recommendation systems and so many other fields of information and technology. To solve these issues, Novel automatic text classification technique using Natural Language Processing (NLP) is presented in this work. The performance of presented approach is evaluated in terms of accuracy and sensitivity. It will achieve better results compared to earlier approaches.

KEYWORDS: Text Classification, Text mining, Natural Language Processing.

I. INTRODUCTION

With the advent of World Wide Web, amount of data on web increased tremendously. Although, such a huge accumulation of information is valuable and most of this information is texts, it becomes a problem or a challenge for humans to identify the most relevant information or knowledge. Text Classification is the task of classifying a document under a predefined category [4].

Text classification of labeled documents is growing its necessity enormously because there are large amount of documents growing all over the World Wide Web (WWW). Text mining is a research area that deals with the construction of models and patterns from text resources, aiming at solving tasks such as text categorization and clustering, taxonomy construction, and sentiment analysis.

Automatic text classification has always been an important application and research topic since the inception of digital documents. Today, text classification is a necessity due to the very large amount of text documents that we have to deal with daily. In general, text classification includes topic based text classification and text genre-based classification. Topic-based text categorization classifies documents according to their topics. Texts can also be written in many genres, for instance: scientific articles, news reports, movie reviews, and advertisements. Genre is defined on the way a text was created, the way it was edited, the register of language it uses, and the kind of audience to whom it is addressed.

A document collection is any grouping of text documents that can be used in text analytics. Even though the size of a collection may vary from a few to millions of documents, from the text analysis perspective, more is better. Text documents can be classified through various kinds of classifiers. Labeled text documents are used to classify the text in supervised classifications. Assigning categories of documents, which can be a web page, library book, media articles,

ISSN PRINT 2319 1775 Online 2320 7876
Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

gallery etc. has many applications like spam filtering, email routing, sentiment analysis etc. We would like to demonstrate how we can do text classification using the most common python machine learning and natural language processing packages like: Pandas, Scikit-learn, Numpy and little bit of NLTK.

Classification can be manual or automated. Unlike manual classification, which consumes time and requires high accuracy, Automated Text Classification makes the classification process fast and more efficient since it automatically categorizes document. [6]. This research area, also known as text data mining or text analytics, is usually considered as a subfield of data mining (DM) research, but can be viewed also more generally as a multidisciplinary field drawing its techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR), information extraction (IE) and knowledge management [1].

With NLP the text data can be analyzed to mine the sentiment. NLP (Natural Language Processing) is a technology that enables computers to understand human languages. Deep-level grammatical and semantic analysis usually uses words as the basic unit, and word segmentation is usually the primary task of NLP [2]. NLP (Natural Language Processing) is one of the key technologies for realizing human-computer interaction and artificial intelligence. It is listed as the three major elements of artificial intelligence research with voice processing and image processing. NLP can be defined broadly as a set of methods for processing and analyzing textual data with computers. In the context of radiology, NLP is most often applied to the text from radiology reports, although the general principles can be applied to other textual data extracted from the electronic medical record. In practice, NLP involves a series of steps to convert free-form textual data into structured numeric data that machine learning algorithms can analyze. Potential applications for NLP analysis of radiology reports include information extraction and report classification, machine translation (eg, lay summary generation), and automated impression generation.

NLP has gained much interest in recent years, mostly in the field of text analytics, Classification is one of the major task in text mining and can be performed using different algorithms [3]. NLP technology's unique machine translation and text sentiment analysis functions can prevent people from experiencing poor language communication when travelling abroad and help artificial intelligence understand people's language better.

II. LITERATURE SURVEY

Hassan Raza, M. Faizan, Ahsan Hamza, Ahmed Mushtaq and Naeem Akhtar et. al.,[5] describes Scientific Text Sentiment Analysis using Machine Learning Techniques. They developed a system in which six different machine learning algorithms including Naïve-Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbor (KNN) and Random Forest (RF) are implemented. Then the accuracy of the system is evaluated using different evaluation metrics e.g. F-score and Accuracy score. To improve the system' accuracy additional features selection techniques, such as lemmatization, n-graming, tokenization, and stop word removal are applied and found that our system provided significant performance in every case compared to the base system.

P. M. ee, S. Santra, S. Bhowmick, A. Paul, P. Chatterjee and A. Deyasi et. al., [7] describes Development of GUI for Text-to-Speech Recognition using Natural Language Processing In the present paper, a Text-to-speech synthesizer is developed

ISSN PRINT 2319 1775 Online 2320 7876
Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

that converts text into spoken word, by analysing and processing it using Natural Language Processing (NLP) and then using Digital Signal Processing (DSP) technology to convert this processed text into synthesized speech representation of the text. Here we developed a useful text-to-speech synthesizer in the form of a simple application that converts inputted text into synthesized speech and reads out to the user which can then be saved as an mp3 file.

Matic Perovšek, Janez Kranjc, Tomaž Erjavec, Bojan Cestnik, Nada Lavra et. al., [8] describes Text-Flows: A visual programming platform for text mining and natural language processing. This analysis presents TextFlows, a web-based text mining and natural language processing platform supporting workflow construction, sharing and execution. The platform enables visual construction of text mining workflows through a web browser, and the execution of the constructed workflows on a processing cloud. This makes TextFlows an adaptable infrastructure for the construction and sharing of text processing workflows, which can be reused in various applications.

M. Ali, S. Khalid, M. I. Rana, and F. Azhar et. al., [9] presents A probabilistic framework for short text classification. a probabilistic framework for short text classification. Proposed classification model is composed of three major modules i.e. pre-processing of unstructured text, learning of probabilistic model and the classification of unseen data by using learned model. This framework is trained and tested by using news headlines dataset containing six different news categories i.e. politics, sports, business, weather, showbiz and terrorist. During the experimental evaluation has achieved good classification accuracy by using bigram as feature which demonstrates the effectiveness of described short text classification approach.

S. Z. Mishu and S. M. Rafiuddin et. al., [10] describes Performance analysis of supervised machine learning algorithms for text classification. This approach applied these classifiers on different kinds of labeled documents and measures the accuracy of the classifiers. An Artificial Neural Network (ANN) model using Back Propagation Network (BPN) is used with several other models to create an independent platform for labeled and supervised text classification process. An existing benchmark approach is used to analysis the performance of classification using labeled documents. Experimental analysis on real data reveals which model works well in terms of classification accuracy.

III. NOVEL AUTOMATIC TEXT CLASSIFICATION TECHNIQUE

In this work, novel automatic classification technique using Natural Language Processing is presented. The workflow diagram of presented approach is shown in Fig. 1.

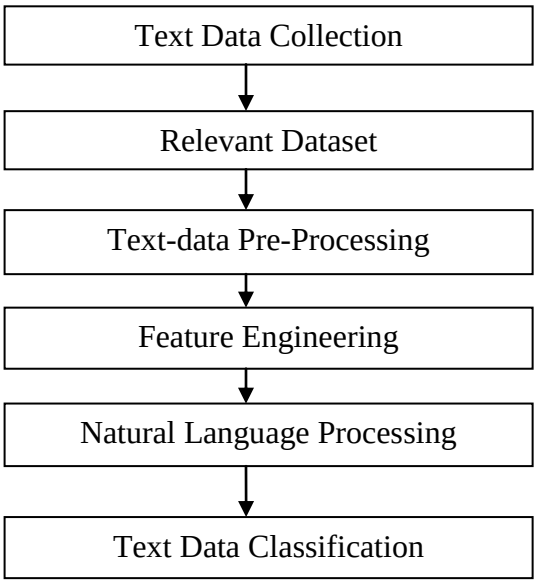

Fig. 1: Work Flow Diagram of Novel Automatic Text Classification

ISSN PRINT 2319 1775 Online 2320 7876
Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

Text Classification is the task of assigning a sentence or document an appropriate category. The categories depend on the chosen dataset and can range from topics. Text Classification problems include emotion classification, news classification, citation intent classification, among others. Text classification datasets are used to categorize natural language texts according to content. The Google Books Ngram dataset is used. This Google Books Ngram Dataset contains frequencies of any set of search strings using a yearly count of n-grams found in sources printed between 1500 and 2012 in Google's text corpora. Languages include English, Chinese (simplified), French, German, Hebrew, Italian, Russian, and Spanish. This data was acquired from Google Books store. Google API was used to acquire the data. Nine features were gathered for each book in the data set.

The column names mostly are self explanatory nevertheless, it will be explained below: i) Title: the title of the book; ii) Authors: name of the authors of the books (might include more than one author; iii) language : the language of the book; iv) generes\categories: the categories associated with the book (by Google store); v) rating\averageRating : the average rating of each book out of 5; vi) maturityRating : wheather the content of the book is for mature or NOT MATURE audience; vii) publisher : the name of the publisher; viii) publishedDate : when the book was published; ix) pageCount: number of pages of the books; x) voters: the number of voters to the book; xi)ISBN : the unique identifier for each book; xii) description: brief introductory description of the book; xiii)price : price of the book on the google books store; ix) currency : the currency of the price in the google books store.

The text is unstructured so it needed to be refined such that machine learning can be done. A standard collection of pre-processing techniques is listed below, together with sets of functionalities implemented in our platform: Tokenization: In tokenization, meaningful tokens are identified in the character stream of the document, such as words or terms. Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, certain characters, such as punctuation is filtered out during the process.

Stopwords, symbols, Url's, links are removed such that classification can be achieved with better accuracy. The features like stop words & punctuation removal, lemmatizationnm, etc.) along with n-grams and then again compute the accuracies. The latter approach helps to reduce the noise and complexity of the data. Lower case and punctuation removal: This step transforms the text into lower case which reduces variation of same word, e.g., after transformation 'Employee' and 'employee' are treated as the same word. Punctuations increase the size of training data and usually do not contribute much to text analysis, thus are removed. Stemming and lemmatization: In a document, same word can be expressed in different forms, e.g. 'kill', 'kills', 'killing'. Moreover, words can be represented in different syntactic categories that have the same root form and are semantically related, e.g. 'irony', 'ironic.

Stopwords removal: Stopwords are extremely common words which are of little value in helping select documents and such words are excluded. Some published stop-words lists are available for example in Snowball stop word list published with the Snowball Stemmer and Terrier stop word list published with the terrier package. However, stop-words of different domains are different. For medical domain, words like 'pill', 'patient' occur in most documents and such words are

ISSN PRINT 2319 1775 Online 2320 7876
Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

considered stop-words while for computer product domain, potential stop-words list consists words such as 'CPU', 'memory', etc. Generally, common stop-words list does not cover such terms, a domain specific stop-words list can be complied base on acquired domain knowledge.

N-grams: N-grams of texts are a set of co-occurring words within a given window size n, i.e. window size of unigrams, bigrams trigrams is one, two, three respectively. An example for the sentence 'he is a girl', unigrams are 'she', 'is', 'a', 'girl' while bigrams are 'she is', 'is a', 'a girl' and trigrams are 'she is a', 'is a girl.

Various features are extracted as per the semantics and are converted into probabilistic values. TF//IDF technique is used for extracting relevant features. Bag of words was also taken into consideration, unigrams, bigrams were also extracted. Relevant features are extracted by which the classification can be achieved. TF-IDF stands for term frequency-inverse document frequency and it is a measure, used in the fields of information retrieval (IR) and machine learning, that can quantify the importance or relevance of string representations (words, phrases, lemmas, etc) in a document amongst a collection of documents (also known as a corpus).

Term frequency works by looking at the frequency of a particular term you are concerned with relative to the document. There are multiple measures, or ways, of defining frequency: i) Number of times the word appears in a document (raw count), ii) Term frequency adjusted for the length of the document (raw count of occurences divided by number of words in the document), iii) Logarithmically scaled frequency (e.g. log(1 + raw count)) iv) Boolean frequency (e.g. 1 if the term occurs, or 0 if the term does not occur, in the document).

Inverse document frequency looks at how common (or uncommon) a word is amongst the corpus. IDF is calculated as follows where t is the term (word) we are looking to measure the commonness of and N is the number of documents (d) in the corpus (D). The denominator is simply the number of documents in which the term**,** t**,** appears in

$$\mathrm{idf}(t, D) = \log\left(\frac{N}{\mathrm{count}(d \in D : t \in d)}\right) \quad (1)$$

NLP enables computers to understand natural language as humans do. Whether the language is spoken or written, natural language processing uses artificial intelligence to take real-world input, process it, and make sense of it in a way a computer can understand. Just as humans have different sensors such as ears to hear and eyes to see computers have programs to read and microphones to collect audio. And just as humans have a brain to process that input, computers have a program to process their respective inputs. At some point in processing, the input is converted to code that the computer can understand.

Natural language processing (NLP) involves the techniques of multiple areas in artificial intelligence, computational linguistics, mathematics and information science, it the approach to make computer understand natural language and perform certain tasks. NLP can be utilized to analyze semantic and grammatical sutures of text while such analysis cannot be performed by text mining.

There are many different natural language processing algorithms, but two main types are commonly used. Rules-based system: This system uses carefully designed linguistic rules. This approach was used early on in the development of natural language processing, and is still used. Machine learning-based system: Machine learning algorithms use statistical methods. They learn to perform tasks

ISSN PRINT 2319 1775 Online 2320 7876
Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

based on training data they are fed, and adjust their methods as more data is processed. Using a combination of machine learning, deep learning and neural networks, natural language processing algorithms hone their own rules through repeated processing and learning.

Document classification (also called text categorization) refers to automated assigning of predefined categories to natural language texts. A primary application of text categorization systems is to assign subject categories to documents to support information retrieval, or to aid human indexers in assigning such categories. Text categorization components are also increasingly being used in natural language processing systems for data extraction. Finally the text is classified into categories of documents, which can be a web page, library book, media articles, gallery etc.

IV. RESULT ANALYSIS

In this section, novel automatic classification technique using Natural Language Processing is presented. The result analysis of presented approach is evaluated here in terms of classification Accuracy and Sensitivity.

Accuracy: Accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training, data. It is defined as the ratio of number of correctly classified instances to the total number of detected instances.

Sensitivity: It is also known as True Positive Rate (TPR) or Recall. It is measured as the ratio of number of correctly detected positive instances to the total positive instances.

The table 1 shows the performance evaluation.

Table1: Performance Evaluation

Metrics/Methods	Sensitivity (%)	Accuracy (%)
Image based text classification	82.34	84.56
Presented novel automatic classification technique using Natural Language Processing	92.45	94.23

Compared to earlier image based text classification approach; presented approach has better accuracy and sensitivity.

The fig. 2 shows the sensitivity and accuracy comparison graph.

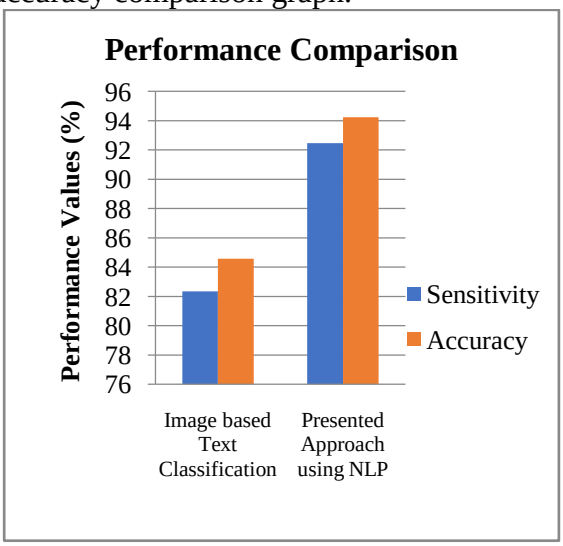

Fig. 2: Performance Comparison Graph

The Fig. 2 shows the performance comparison graph where x-axis represents different text classification techniques and y-axis represents performance values in terms of percentage. Presented novel automatic classification technique using Natural Language Processing has better accuracy and sensitivity than image based text classification approach.

V. CONCLUSION

In this work, novel automatic classification technique using Natural Language Processing is presented. The Google Books Ngram dataset is used. The

ISSN PRINT 2319 1775 Online 2320 7876
Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

collected data is preprocessed. Various steps are being followed in the pre-processing phase; the text is being cleaned by removing unnecessary text. Punctuation and lemmatization are being done such that the data is refined in a better way. TF//IDF technique is used for extracting relevant features. Bag of words is also taken into consideration, unigrams, bigrams are also extracted. NLP is utilized to analyze semantic and grammatical sutures of text and to classify the text documents. It is classified as web page, library book, media articles, gallery. The performance of presented approach is evaluated in terms of accuracy and sensitivity. Compared to earlier approaches, presented approach has better performance in terms of accuracy and sensitivity.

VI. REFERENCES

[1] Xi. Peng, "A Comparative Study of Neural Network for Text Classification," 2020 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), *Shenyang*, China, 2020, pp. 214-218, doi: 10.1109/TOCS50858.2020.9339702.

[2] D. Wang, J. Su and H. Yu, "Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language," in *IEEE Access*, vol. 8, pp. 46335-46345, 2020, doi: 10.1109/ACCESS.2020.2974101.

[3] Akib Mohi Ud Din Khanday, Syed Tanzeel Rabani, Qamar Rayees Khan, Nusrat Rouf, Masarat Mohi Ud Din, "Machine learning based approaches for detecting COVID-19 using clinical text data", Int. j. inf. tecnol. (September 2020) 12(3):731–739, doi:10.1007/s41870-020-00495-9

[4] R Ravi Kumar, M Babu Reddy, P Praveen, "Text Classification Performance Analysis on Machine Learning", International Journal of Advanced Science and Technology Vol. 28, No. 20, (2019), pp. 691 – 697

[5] Hassan Raza, M. Faizan, Ahsan Hamza, Ahmed Mushtaq and Naeem Akhtar, "Scientific Text Sentiment Analysis using Machine Learning Techniques" International Journal of Advanced Computer Science and Applications(IJACSA), 10(12), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0101222

[6] N I Widiastuti, Convolution Neural Network for Text Mining and Natural Language Processing", INCITEST 2019, Materilas Science and Engineering, doi:10.1088/1757-899X/662/5/052010

[7] P. M. ee, S. Santra, S. Bhowmick, A. Paul, P. Chatterjee and A. Deyasi, "Development of GUI for Text-to-Speech Recognition using Natural Language Processing," 2018 2nd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), Kolkata, India, 2018, pp. 1-4, doi: 10.1109/IEMENTECH.2018.8465238.

[8] Matic Perovšek, Janez Kranjc, Tomaž Erjavec, Bojan Cestnik, Nada Lavra, "Text-Flows: A visual programming platform for text mining and natural language processing", Science of Computer Programming 121 (2016) 128–152, doi: 10.1016/j.scico.2016.01.001

[9] M. Ali, S. Khalid, M. I. Rana, and F. Azhar, "A probabilistic framework for short text classification," in Proceedings of the IEEE Eighth Annu Comput Commun Work Conf CCWC, 2018, pp. 742–747, Las Vegas, NV, USA, February 2018

[10] S. Z. Mishu and S. M. Rafiuddin, "Performance analysis of supervised machine learning algorithms for text classification," 2016 19th International Conference on Computer and Information Technology (ICCIT), *Dhaka, Bangladesh*, 2016, pp. 409-413, doi: 10.1109/ICCITECHN.2016.7860233.