iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** V **Month of publication:** May 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Natural Language Processing on Big Data

Jaydeep Parmar[1], Prof. Krunal Vaghela[2]
[1]M. Tech, Department of Computer Engineering, School of Engineering RK University, Rajkot, India.
[2]Head of Department (CE, IT, BCA, MCA), School of Engineering RK University, Rajkot, India.

Abstract: ***Natural Language Processing and Big data are both different concept. Big data basically means set of data which are quit large and complex that it's inadequate to process traditional data. Natural Language processing is nothing but the sub part of AI, computational linguistics which is interaction between computer and human languages. It includes spoken words, emotions, and sentiments of human. Natural Language Processing is a process of language that is used by humans. Big data is the concept of data that is large in size, hard to handle, hard to analyse, and it has different types like structured, unstructured, semi structured data. Now, when it combine this both, it can process a language with help of big data concept. Machine can read, write, speak this is concept of NLP. So by processing natural language on machine and by analysing natural language data, decision making process can be done easily. And can use it in politics, healthcare, finance, marketing, etc.***
Keywords: NLP, big data, sentiment analysis, Natural Language Processing, Data analysis C.

I. INTRODUCTION

Big data basically means set of data which are quit large and complex that it's inadequate to process traditional data. The big data use for prediction and analysis process on data, human behaviour or any kind of other data analysis method. That gives us a values from data. Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges that we face with dbms tools and other technologies is capture, curation, storage, search, sharing, transfer, analysis, and visualization.

A. Problem Definition

Natural language Processing is a combination of artificial intelligence and computer Science, which works as a bridge between Machine and Human. Natural language Processing is interaction between human and Machine. And Big data is large amount of data that needs to be analyse for decision making process. So we have to build a bridge between NLP and Big data. And the reason is the development and improvement in technology in today`s world and nearer future. Due to research and improvement in technologies and as per our requirement it's necessary to reduce the gap between machine and human. So that machine can understand Big data and so we can use NLP on Big data to do different task.

B. Introduction to NLP: [1]

Natural Language processing is nothing but the subpart of AI, computational linguistics which is interaction between computer and human languages. It includes spoken words, emotions, and sentiments of human. Natural Language Processing is a process of language that is used by humans .Big data is the concept of data that is large in size, hard to handle, hard to analyse, and it has different types like structured, unstructured, semi structured data. Now, when we combine this both, we can process a language with help of big data concept. Machine can read, write, speak this is concept of NLP. So by processing natural language on machine and by analysing natural language data, decision making process can be done easily. And we can use it in politics, healthcare, finance, marketing ,etc.IN past, NLP were based on machine learning algorithms that is used for natural language processing .part-of-speech tagging POS has used for NLP, and increased by time , research has focused on natural language processing for decision making process can be done easily by help of NLP analytics.

C. Why nlp and big data? [1][2]

Data are both different concept. Big data basically means set of data which are quit large and complex that it's inadequate to process traditional data. Natural Language processing is nothing but the subpart of AI, computational linguistics which is interaction between computer and human languages. It includes spoken words, emotions, and sentiments of human. Natural Language Processing is a process of language that is used by humans. Big data is the concept of data that is large in size, hard to handle, hard to analyse, and it has different types like structured, unstructured, semi structured data. Now, when we combine this both, we can process a language with help of big data concept. Machine can read, write, speak this is concept of NLP. So by processing natural

language on machine and by analysing natural language data, decision making process can be done easily. And it can use it in politics, healthcare, finance, marketing, etc.

II. RELATED WORK

A. Big data for Natural Language Processing: A Streaming Approach [3]

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. And their decisions make a huge impact or make a big difference for any organization.so that there is a race between professional decision maker for their accuracy, efficiency, liability, cost, and other parameters. Processing a huge amount of data has become a major challenge in NLP. At present around 80percent of data are unstructured. Such as web pages, news articles, any organizations data, there are many tools are available for processing such type of data. This tools perform operations on this textual data. And in timely manner it process millions of documents.

B. Understanding pending issues of society and sentiment analysis using social media [4]

Social media such as twitter is nowadays platform for expressing thoughts and experience on trending topic on which millions of user share their idea, there for twitter serves as source of information for making decision and analysis of particular decision. Computer technology has influence are alive to many changes like, live coverage of any of particular event happening, storage, management, and any kind of printing. Twitter also helps to discovering new events and follow stream data model. Large number of data is created every day, so storage of this data is quit difficult to be stored in only small amount of information can be stored. Speed of data entry is very high, Data distribution and production changes over time.

III.PROPOSED FRAMEWORK

This research was focused on to find out the current sentimental trend when we are using Natural Language Processing On BIGDATA and how it can be helpful in business, social network, finance, politics and how sentimental analysis can be done on different data, and how this analysis can help for decision making process. Sentiment analysis is must for any organization, government, company, because by sentiment analysis, the decision making process became easy. I will do sentiment analysis on dataset by using Rapid miner tool. It will help us to and the emotions sentiment of people, and trending topics. Big data and Natural Language Processing is now trending topics of IT Company and it is the future of Data processing. So by combining them we can direct data analysis process in a new way.

IV.EXPERIMENTAL EVALUATION

As noted, until recently most of the interest about unstructured data has focused on Text Mining, Information Retrieval and topic classification tasks. In particular, textual data provided by Web users represents one of the most useful and interesting sources of unstructured information that can be currently found. Under this light, the entire field of Natural Language Processing evolves together with the increase of computational power and the discovery of new techniques which are aimed to interpret such textual information. In order to decide which kind of support should be incorporated into NLTK, a review of the range of existing approaches and corpora for Sentiment Analysis is required. This review should provide considerations in order to decide which elements will receive support during the implementation phase. Useful parameters would take into consideration aspects such as the maturity of proposed approaches, the availability of related resources, licensing terms and issues, and specific implementation requirements that could eventually hinder the feasibility of implementing the specific approach within NLTK. The most interesting results in the field of Sentiment Analysis are mainly obtained employing Machine Learning techniques or lexicon-based analyses. Machine Learning can be described as the process of automatically inferring patterns and structures from data, ideally providing as few as possible domain-specific instructions to the machine that has to accomplish the task. As we will see, it is still not possible to simply ask a machine to guess a pattern without providing it with some guidelines that react our assumptions: this extreme edibility is something that still has to be reached, especially if we take performance issues into consideration. However, a reasonable balance between this idea and very specific step-by-step instructions can still be reached thanks to Machine Learning and statistical methods.

A. Pre-processing and feature extraction

Sentiment Analysis is closely bound to Information Retrieval and Natural Language Processing tasks. However, treating an opinion mining process using the same structures and techniques employed in other text classi_cation tasks usually produces unexpected and disappointing results. Working with text, we need to use features that are only based on words. In order to achieve good results, however, we need to make assumptions about the words that will yield more benefits to our work; this is what is usually described as feature selection in most Natural Language and Machine Learning approaches, and the difference with other text classi_cation

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

tasks mostly resides in this process. The words we care about in Sentiment Analysis are dfferent from those that we would usually employ in other tasks as topic classi_cation or named entity recognition; in our specific environment, for example, adverbs and adjectives are more interesting and useful than nouns and verbs, since they usually convey more information about sentiments and subjectivity.

Moreover, what we are describing as word should be better called token; what this term generalization is supposed to convey is the fact that, especially in Sentiment Analysis, the data components that can help our task are not only restricted to regular words. In fact, when interpreting the sentiment of a text, we are usually intended by other symbols: punctuation marks, uppercase letters and also groups of characters that form so-called emoticons. With this in mind, we can now analyse several feature extraction strategies that have been used and evaluated in previous research. The first step to perform in order to extract features from raw text consists in splitting strings into smaller constituents (our tokens). Tokenizers are often employed to break long documents into para-graphs, paragraphs into sentences, sentences into individual words and symbols and sometimes, depending on the task, even words can be further split into characters. In our work, we will mostly assume that a token refers to an instance found at word-level, which also includes punctuation symbols, emoticons, and eventually n-grams. Tokens obtained as result of a tokenization process can later be filtered so that the system will only keep the most meaningful ones with regards to the specific application, or they can be pre-processed in several ways depending on the domain we are working into. Different kinds of tokenization algorithms can then be employed to prepare the text for feature extraction.

B. Whitespace Tokenization

Whitespace tokenizers simply split text whenever a space, tab, or newline is encountered. E.g.:

Sentence: “Whose woods these are I think I know.”

Tokens: ['Whose', 'woods', 'these', 'are', 'I', 'think', 'I', 'know.']

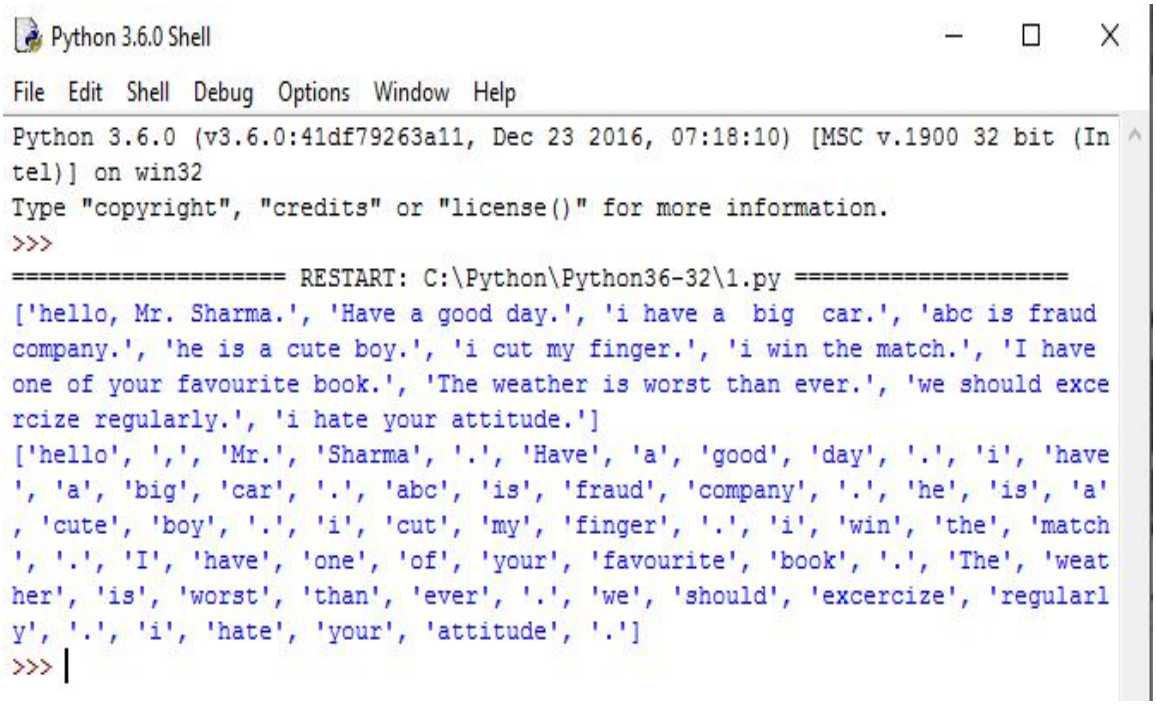


Fig.1 Whitespace Tokenization

C. Stop Words

In many applications word features are added based on their frequency counts. While this may produce good results in general, especially from a performance point of view, it has often been shown that removing some of the most frequent words can generally improve classi_cation results with Machine Learning methods. These common words are usually called stop words, and their removal is based on the assumption that tokens that appear too frequently in all documents are likely to be too generic and thus not very informative.

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

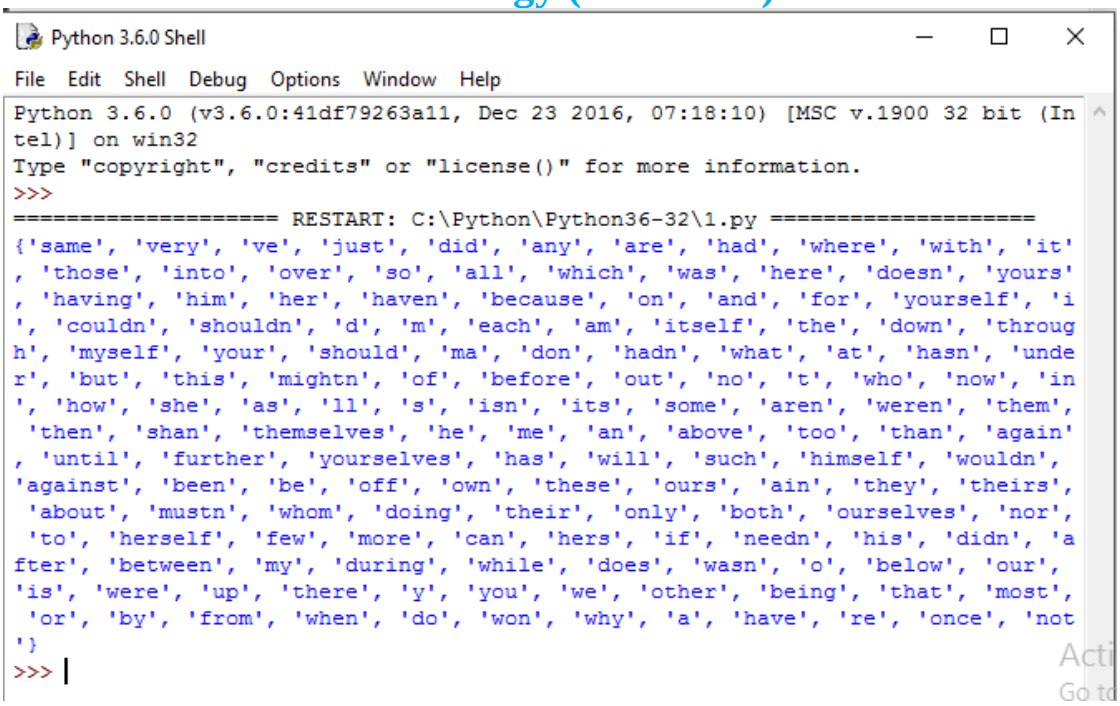


Fig.2 Displaying Stop-words

D. Stemming:

In those cases where a more compact feature space would produce better results, words can be reduced to shorter tokens using a stemmer. Words like “riding”,” ride “and “rides “could be for example reduced to the same string ?love?; while this opera option reduces the sparseness of the data, it decreases at the same time the specificity of each token. In the given example, we can see that we lost the suffixes that could have helped us to discriminate between the lover and the loved, which is somewhat ungraceful especially for Sentiment Analysis.

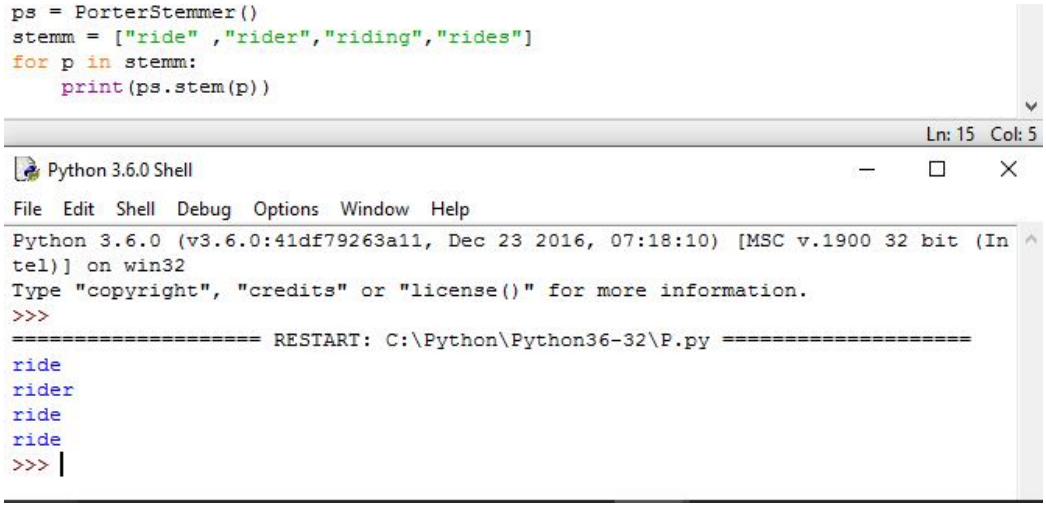


Fig.3 Stemming of words

E. Implementation

For particular paragraph we have to implement the Natural Language Processing. For this first of all we have to decide what should analyse. Take an example of below paragraph. That paragraph should analyse. Example text: “hello, Mr. Sharma. Have a good day. I have a big car. Abc is fraud company. He is a cute boy. I cut my finger. I win the match. I have one of your favourite book. The weather is worse than ever. We should exercise regularly. I hate your attitude.”

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

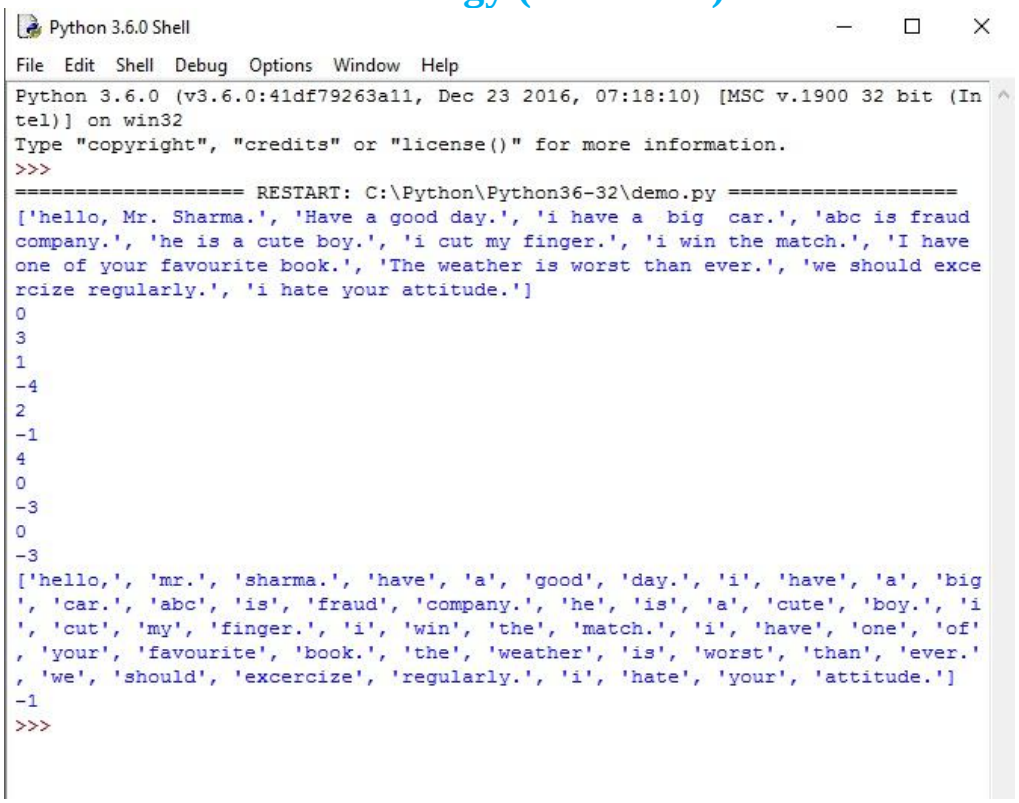


Fig. 4 Sentiment analysis of paragraph

Now, when NLP applied on this paragraph with help of the python coding then. At into sentence and word tokenize breaks the paragraphs into words. And each word represents one token in word tokenize. And each sentence represents one token in sent-tokenize. Now, NLP will analyse the each sentence and doing sentiment analysis on each sentence. This sentiment analysis could be done on the basis of sentiment level of the sentence. For measuring this sentiment it has dictionary that can measure the sentiments of words. So in this words are analysed on the basis of the intensity level of sentiments. For this when we are doing sentiment analysis, it will analyse the word in sentence. As per word's intensity of sentiment it will give it some number on basis of intensity between -5 to +5. If words represents the negative sentiment or emotions then it will scale less negative to more negative in -1 to -5 and neutral sentiment or emotion represents the 0 number. Positive sentiment or emotions can be represents into scale.

V. ISSUES AND LIMITATION

There are some issues which should understand first. Language problem is main problem because it supports only English language in NLTK when we are working with python it does not support other language analysis. Other issue is the storage of data. When we talks about big data. At that time the first thing comes in our mind is size. So due to large amount data the storage of data is the issue. We have to organize storage which is large in capacity of data. When we are analysing the streaming data at that time we do not have to need big storage but we should have high bandwidth so that large amount of data can be analysed flawlessly. When we are doing Natural Language processing using NLTK with help of python language. at that time when we have to analyse the data which is in large size or Big data . Then we should have the system that has more processing power and more RAM. So that we can analyse the data without interruption and time consumes less. And so that efficiency would be greater.

VI. CONCLUSION

Natural Language Processing and Big data are both different concept. Big data basically means set of data which are quit large and complex that it’s inadequate to process traditional data. Natural Language processing is nothing but the subpart of AI, computational linguistics which is interaction between computer and human languages. It includes spoken words, emotions, and sentiments of human. Natural Language Processing is a process of language that is used by humans. Big data is the concept of data that is large in size, hard to handle, hard to analyse, and it has different types like structured, unstructured, semi structured data. Now, when we combine this both, we can process a language with help of big data concept. Machine can read, write, speak this is concept of NLP.

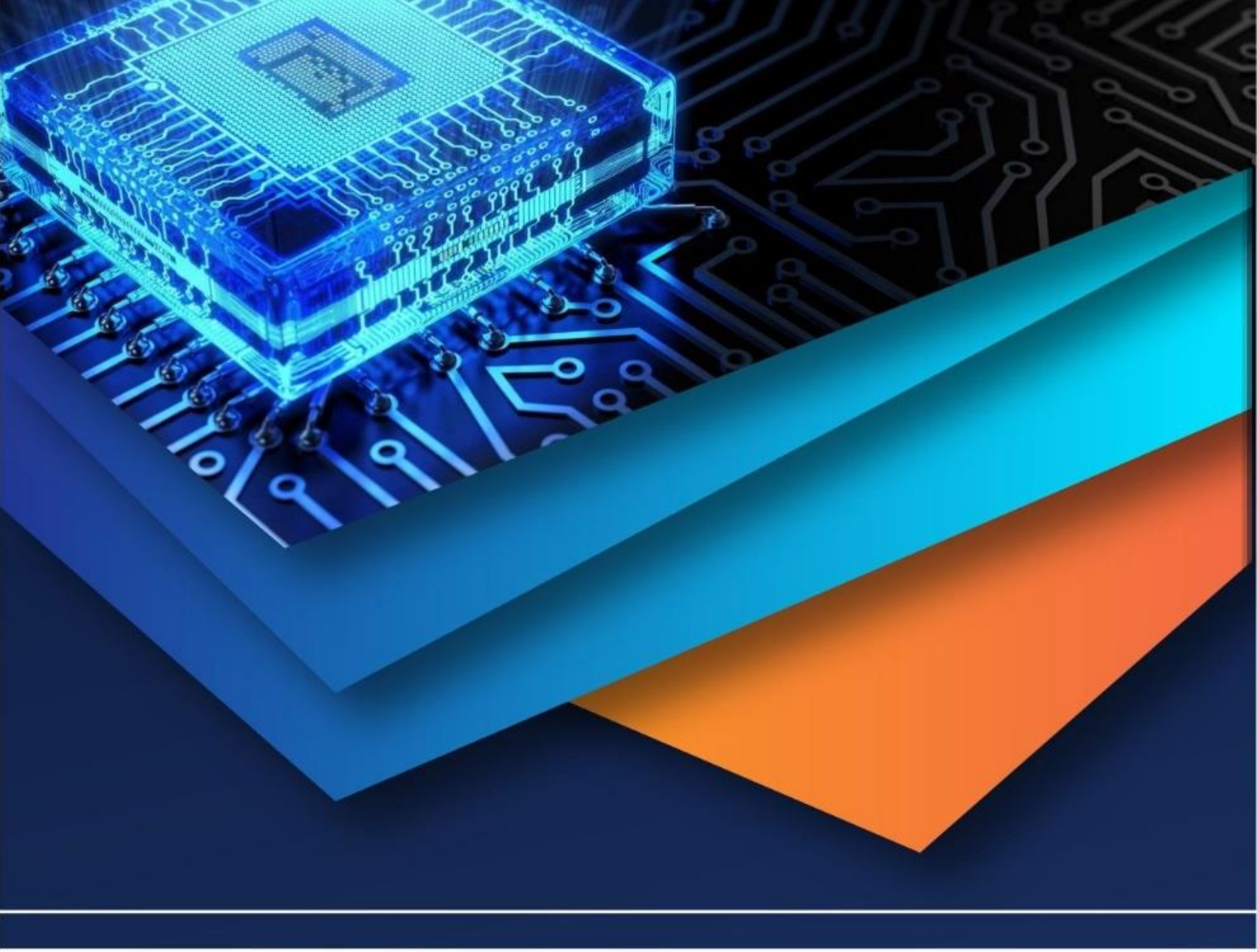
So by processing natural language on machine and by analysing natural language data, decision making process can be done easily. And it can use it in politics, healthcare, finance, marketing, etc.

VII. ACKNOWLEDGEMENT

We thank our institute School of Engineering RK University for providing Required Knowledge. We would also like to thank Vice Chancellor of RK University, Director of School of Engineering RK University and all the people who helped us to make this possible.

REFERENCES

[1] http://www.doc.ic.ac.uk/ nd/surprise 97/journal/vol1/hks/
[2] http://www.dataversity.net/natural-language-processing-big-data-powerful-combination
[3] Rodrigo Agerri, Xabier Artola, Zuhaitz Beloki, German Rigau, Aitor Soroa, “Big data for Natural Language Processing: A streaming approach “, R. Agerri etal. / Knowledge-Based Systems 79 (2015) 36?42
[4] Jong-Seon Jang, Byoung-In Lee, Chi-Hwan Choi, Jin-Hyuk Kim, Dong-Myung Seo, Wan-Sup Cho, “Understanding Pending Issue of Society and Sentiment Analysis Using Social Media”, Information Technology Research Center,korea-2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY