IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Transcript Based Plain Text Analysis Using Natural Language Processing

[1]Abhay R. Chaskar, [2]Shraddha D. Gadekar, [3]Shiva S. Saran, [4]Ballaleshwar R. Raut, [5]Anjali Almale

[1,2,3,4]Software Engineer, [5]Professor Comp. department

[1,2,3,4,5]Computer Department, Pune, India

Savitribai Phule Pune University, Bhivarabai Sawant Institute of Technology and Research, Pune

Abstract: NLP (Natural Language Processing) is an innovation that empowers PCs to comprehend human dialects. The profound level syntactic and semantic examination normally utilizes words as the fundamental unit, also, word division is normally the essential errand of NLP. This paper presents the Analysis of plain text using NLP method and applies the ideas of deep learning, this paper expects to develop such a framework utilizing NLP by giving organized normal conversation questions, keywords, parameters, or any sentence as information, and after the identification of these inputs getting chat ID, chat Type, stored in the database so any user who needs can access it from the database. The steps involved in this process or technique are tokenization, lemmatization, part of speech labeling, parsing, and identification of input. The paper would give a general perspective on the utilization of Natural Language Processing (NLP) and utilization of standard articulations.

Index Terms **- Text Natural Language Processing (NLP), Query Language, Tokenization, Normalization, Lemmatization, Semantic Analysis, POS labeling, Parsing, planning, Regular Expression.**

I. INTRODUCTION:

In this quick innovatively propelling world, it has become significant for people to associate with PCs to give help with numerous fields like medication, instruction, space research, and so forth Recovery of the necessary data from the information base is a dreary interaction. The basic concept of the project is for us to develop and build a machine that will go through the transcript between two peoples and identify the main purpose or issue of the particular conversation. Many businesses, from small startups to large companies opt to outsource processes to Business Process Management Companies. They provide front office BPM tasks commonly, including customer-related services such as tech support, sales, and marketing. These tasks are mostly achieved through the mode of phone calls or live chats. They are processed at contact centers which deal with 1000s of interactions between client's customers and Business reps on daily basis. But most of them don't have a process in place to classify the reason for these interactions which involves manual analysis to build a data model to further analysis. This is a time-consuming and labor-intensive process. So based on the parameters we give to the machine-like payments, complaints, errors, keywords. Based on this our system should process a thousand of these transcripts and develop a data model where we can store these transcript data for further data analysis. To discover an answer for such an issue and work with human communication with PCs, Natural Language Processing (NLP) procedures are utilized. Another significant use of NLP is chatbot (Chat Robot) that can be utilized for voice or literary communications.

II. LITERATURE SURVEY:

In 1972, W.A. Woods fostered a framework that gave a scan interface for the data set framework that put away data about the stone examples that were brought from the moon for research. This framework utilized two data sets, the compound analyses, and the writing references. This framework utilized Augmented Travel Network (ATN) parser and semantics of Woods. This Structure was shown casually at the Second Annual Lunar Science meeting in 1971. Lifer/Ladder framework (1978) was one of the great inquiry interface procedures (for an example NLP system) which utilized semantic language structure for parsing the information or input also, the question created was given as contribution on a conveyed data set framework. This framework upholds single table questions and basic join inquiries in the event of various tables. we proposed a system that provides an interface for the users to pose questions as input in their natural language. The primary goal of this system is to generate or extract the information from the input from the NLP process and store it into a database for further data analysis. This system includes a feature of eliminating spelling errors from user inputs and used the Word Pair Mining Technique for the same. Then the text input in English is mapped or get identified to an equivalent query from the database. The system goes through the tokenization, stemming, syntactic, semantic phases. The user may ask the question in speech format which is then converted to text. The Structural design comprises tokenizing, normalization, segmentation, stemming, lemmatization, parsing, and identification stages. The input natural language query then goes through morphological investigation then, at that point, the semantic investigation is trailed by a POS labeling stage. The input language query is then parsed utilizing the parser and after the recognizable proof of the issue, chat type and chat id is put away in the data set for additional information examination.

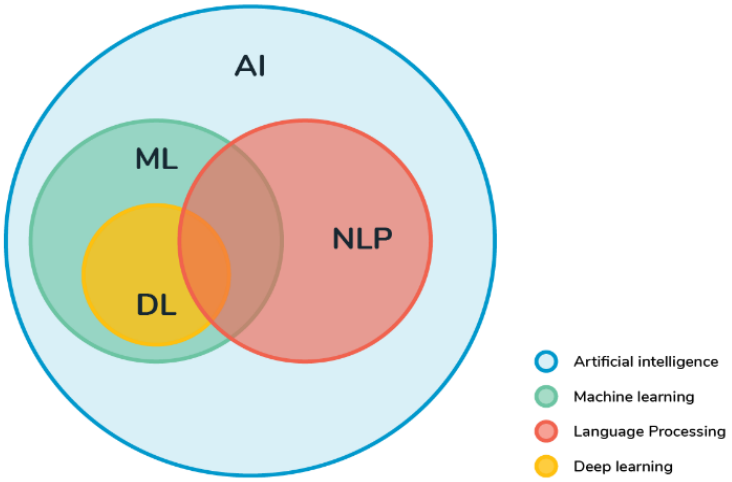

Fig.01

3.1 Existing System:

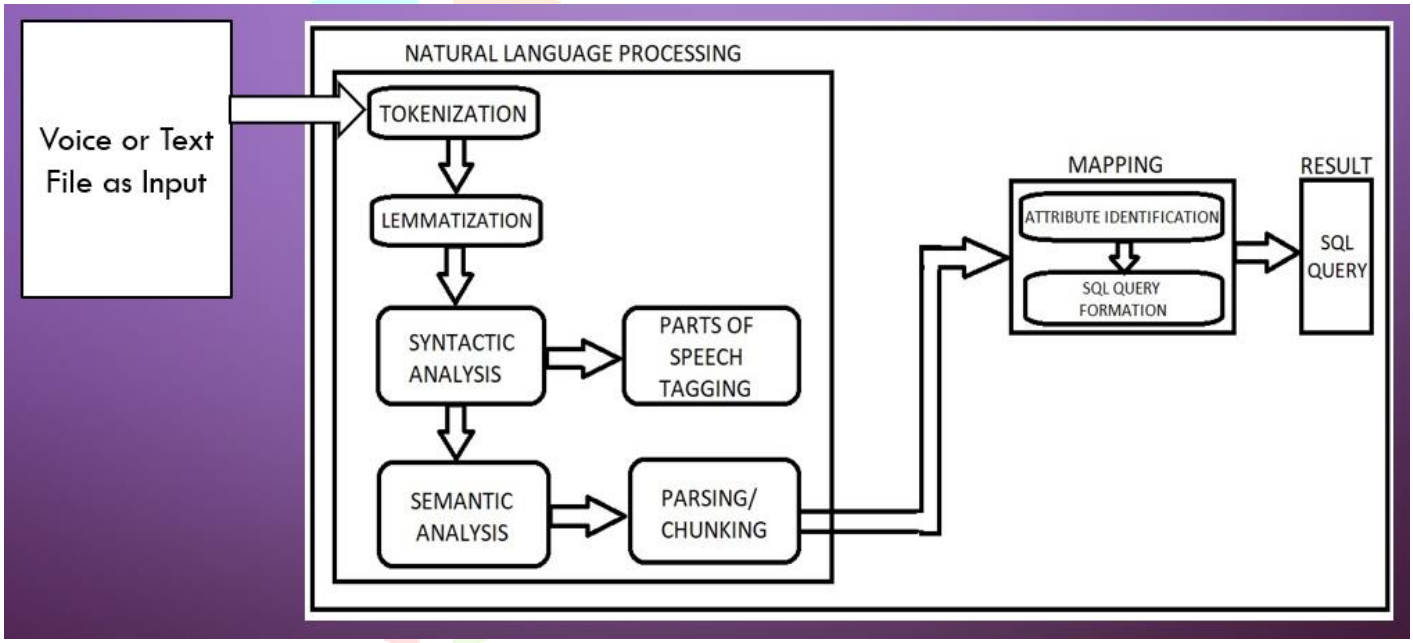

Fig.02 Existing System

3.2 Natural Language Processing:

Natural language Processing assists PCs with speaking with people in their own language and scales other language-related assignments. For instance, NLP makes it workable for PCs to understand text, hear the discourse, decipher it, measure estimation and figure out what parts are significant. Following are the basic steps involved in NLP processing Techniques which we utilized in this framework.

Packages utilized in this project are: Streamlit, Textblob, Spacy, Gensim, Neattext, Matplotlib, Wordcloud.

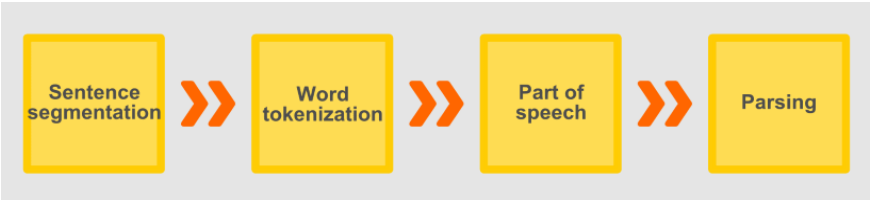

Fig.03 Steps Involved in NLP

3.3 Text Segmentation:

At the point when a PC measures text data, it's nothing but a handling activity of words in the content data. The first issue it faces is the division of words. For each NLP measure, word division/segmentation assumes an essential part. Word segmentation can isolate composed or spoken content into significant word labels, and decide the division range between words in English spaces, so the division between words is self-evident. Then again, there is no space between words in the English text. As of now, the PC should isolate the words in the content to have the option to accurately handle the content substance. It very well may be seen that text division is a significant piece of NLP, particularly when the measure of text data to be handled is huge, the extent of the substance will be more extensive. Right now, there will be numerous expert jargon, how to effectively recognize sectioning text data is a significant factor that influences the presentation of the word division framework. When handling English content data, words are the most fundamental unit. Since there are spaces between English words, the essential technique for word division in English content is presented here:

Word-by-word travelling coordinating technique:

It matches the content substance with the words put away in the word reference one by one until every one of the words in the content substance is portioned out. Since this technique needs to analyze every one of the words in the word reference individually, it sets aside time, the PC is moderate, what's more, the word division productivity isn't high.

Invert maximum matching technique:

It is fundamentally equivalent to the rule of the most extreme coordinating with strategy, the principal distinction is that the examining course is inverse when contrasting, that is: if the examining is from left to right at the point when the most extreme coordinating with technique is thought about, the maximum coordinating with strategy is switched.

3.4 Tokenization:

It is the initial step that is utilized to break a sentence into more modest significant tokens as a rule these are words. In the proposed framework, we applied tokenization when the content information is gotten from the client and the tokens got are put away as a rundown. We have utilized the word tokenize module of the word tokenizes library in Python. Most importantly, understanding the significance of Tokenization, is essentially parting of the entire content into the rundown of tokens, records can be anything like words, sentences, characters, numbers, accentuation, and so forth Tokenization enjoys two fundamental benefits, one is to decrease search with a critical degree, and the second is to be successful in the utilization of extra room.


```
0 : ""Token":0609, "Lemma":0609"
1 : ""Token":162041, "Lemma":162041"
2 :
""Token":AgentFreehand,
"Lemma":AgentFreehand"
3 :
""Token":understand,
"Lemma":understand"
4 : ""Token":Ill, "Lemma":Ill"
5 : ""Token":is, "Lemma":be"
6 : ""Token":Ill, "Lemma":ill"
7 : ""Token":feedback, "Lemma":feedback"
8 : ""Token":recorded, "Lemma":record"
9 :
""Token":complaint, "Lemma":complaint"
10 : ""Token":0609, "Lemma":0609"
11 : ""Token":162054, "Lemma":162054"
```


Fig.04 Tokens & Lemmas

The way toward planning sentences from character to strings and strings into words are at first the fundamental strides of any NLP issue because to see any content or record we need to understand the meaning of the text by interpreting words/sentences present in the text. Tokenization is an indispensable piece of any Information Retrieval (IR) framework, it's difficult to include the pre-interaction of text yet in addition produces tokens separately that are utilized in the ordering/positioning cycle. There are different tokenization' procedures accessible among which Porter's Algorithm is quite possibly the most unmistakable strategy.

3.5 Stemming and Lemmatization:

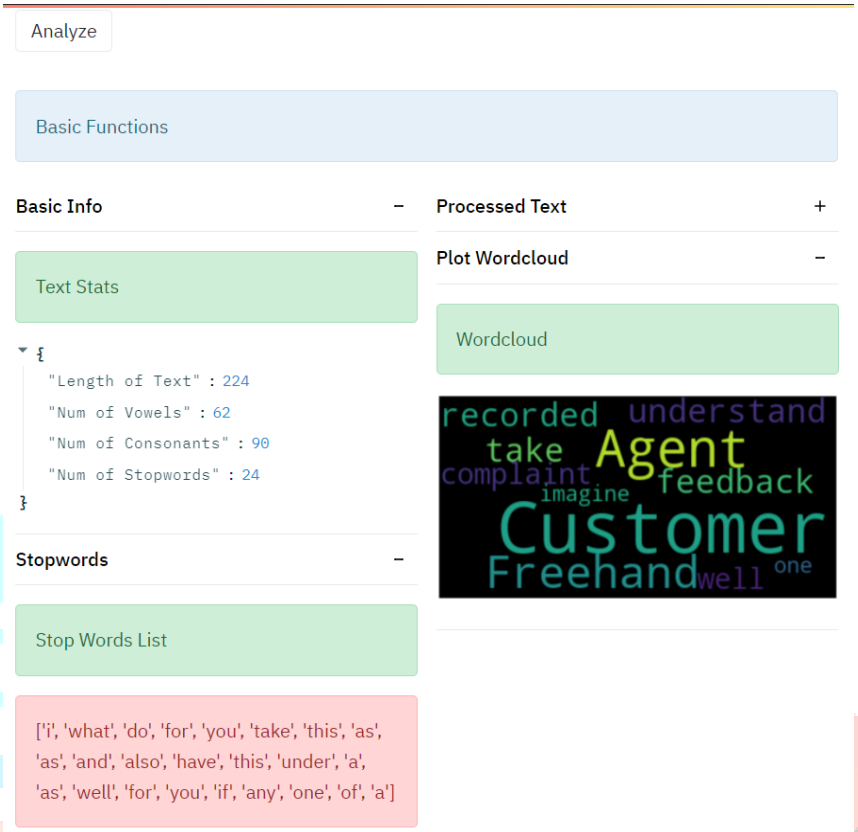

Fig.05 Streamlit UI

The expanding size of information and data on the web is unsurpassed high in recent years. This gigantic information and data request vital devices and strategies to remove deductions without any difficulty. "Stemming is the way toward diminishing bent (or in some cases inferred) words to their promise stem, base or root structure - for the most part, a composed type of the word." For instance, what stemming does, fundamentally it cuts off all the additions. So after applying a stage of stemming on "playing", it becomes "play", or like, "asked" becomes "inquire". This interaction is like stemming where the root words or lemma of every one of these tokens are gotten from the yield of the past advance and are put away in another rundown. Lemmatization is picked over stemming since the way toward stemming doesn't generally demonstrate to be exact since it eliminates the prefix or postfix of a word. While in lemmatization, the roots are coordinated with its lemmas contained in a word reference, and subsequently, more precise results were acquired.

Lemmatization generally alludes to get things done with the legitimate utilization of jargon and morphological investigation of words, regularly meaning to eliminate inflectional endings just and to return the base or word reference type of a word, which is known as the lemma. In straightforward words, Lemmatization manages the lemma of a word that includes lessening the word structure in the wake of understanding the grammatical form (POS) or setting of the word in any archive.

3.6 Syntactical Analysis:

In the syntactic investigation, each of the lemmatized tokens is investigated and as indicated by their unique circumstance of appearance, every token is labeled with a POS. Here, each word and its tag are pressed into a tuple and a rundown of all such tuples is acquired. In the proposed framework, the Stanford POS Tagger is utilized for POS labeling. This tagger is liked over the POS tagger present in the NLTK bundle as it gives more exact labeling.

3.7 Semantic Analysis:

In the semantic examination, we attempt to make feeling of the tokens so the framework could continue with the SQL inquiry development. This is accomplished by the cycle of parsing (or piecing). In the proposed framework, the RegExpParser() (ordinary articulation parser) is utilized for parsing the POS labeled input information. This parser lumps the information dependent on an ordinary articulation. In our work, a normal articulation is outlined such that an expression has the source and objective data is arranged into a different piece and are extricated through a standard-based worldview.

3.8 Part-Of-Speech (POS) Labelling:

After the content is tokenized, you would then be able to break down the piece of the discourse of each word. Grammatical feature labeling alludes to the process of allotting a grammatical form or labeling the jargon type in the text to each word in text data. It's anything but a computer to recognize the grammatical feature of each word in the content. Since there are numerous vocabularies in English writings, and similar words may have various grammatical features in various language passages, and the blends between words are different, the consequences of utilizing the standard technique in English writings may not be ideal. Generally utilized English Part -of-speech labeling is a factual technique. The PC counts all the content data to get the likelihood of mark co-event and the likelihood that the word addresses a specific grammatical form.

4.1 Proposed System:

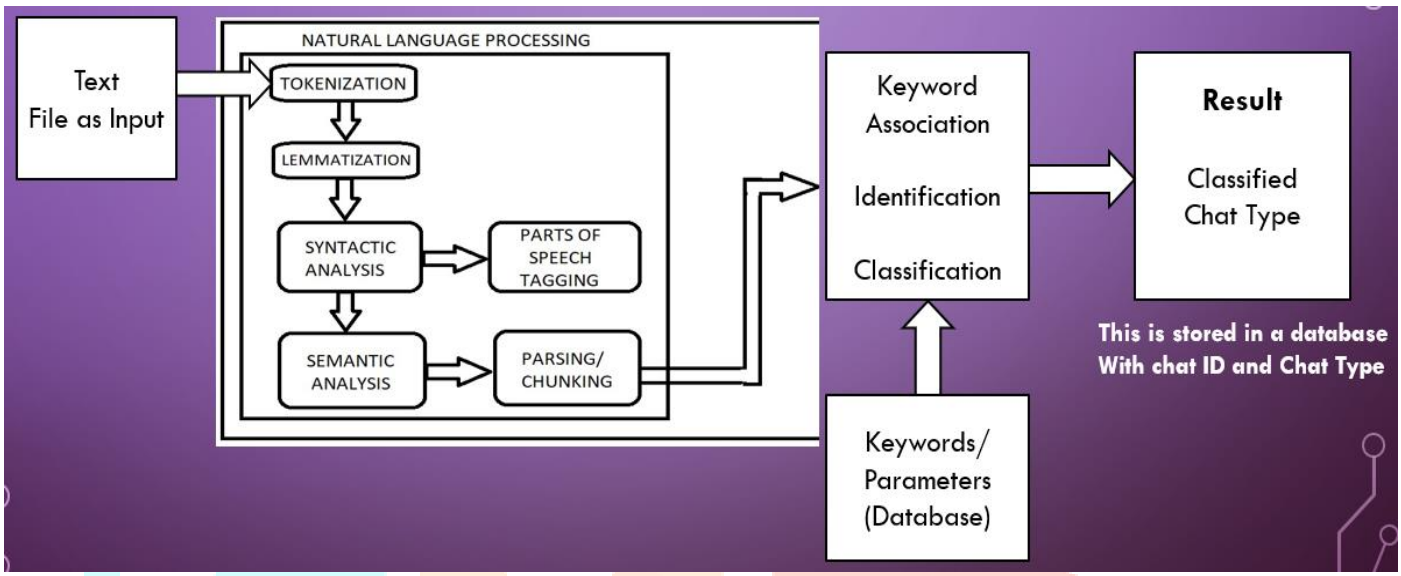

Fig.06 Proposed System

4.2 Architectural Flow:

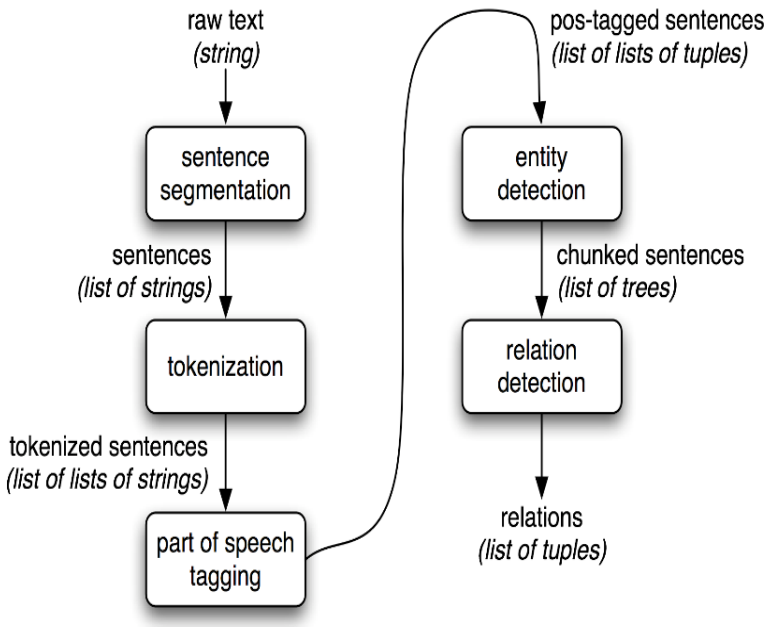

Fig.05 Architecture of NLP Process

4.3 Algorithm:

Step 1: Development of Chat Type from Natural Language Conversation.
Step 2: Input Natural Language query in English text.

Step 3: Tokenize the Conversation between two Users as input into list of words.
Step 4: Lemmatize the list of words.

Step 5: Perform Part of Speech tagging.

Step 6: Parsed sentence = Parse using regular expressions.

Step 7: If Keyword from Database ∈ Parsed sentence.

Step 8: Then Extract them, and classify them into chat type.

Step 9: Classified Chat type (e.g., Payment's type)

Conclusion and Future Scope:

Normal Language Processing has set another norm in doing likewise. This work presents a clear picture of the means that are associated with NLP. Different measures like tokenization, segmentation, lemmatization, syntactic, and semantic analysis or investigation are completed to produce a comparable Chat type just as talk ID from an Input natural language queries. The normal speed is 1.94 occasions that of the word segmentation strategy dependent on BI-LSTM-CRF organization. Based on these two arrangements of information, the crossbreed network word segmentation technique proposed in this paper has great execution in English word division.
In future work, we can consider investigating the effect of various component extraction techniques furthermore, include determination techniques on the model, in this way further upgrading the learning capacity of the model. Following are the future enhancements that can be incorporated; The information got can be of sound structure, which can be changed over into text-based configuration; The SQL question could be of more prominent intricacy; The data set could be bigger in the wording of characteristics and tuples. Additionally, there could be numerous tables of related information which can be gotten to utilizing JOIN catchphrase; It very well may be utilized to make chatbots for different areas which handle huge data sets and can assist clients to get to them with more prominent facilitate; The yield could be changed over as a sentence then, at that point into a sound arrangement to make the framework more intelligent; This work can likewise be reached out to different dialects.

REFERENCES

[1] M. E. Tibbo, C. C. Wyles, S. Fu, S. Sohn, D. G. Lewallen, D. J. Berry, and H. Maradit Kremers, "Use of regular language preparing devices to identify and characterize periprosthetic femur breaks," J. Arthroplasty, vol. 34, pp. 2216–2219, Oct. 2020.

[2] M. Joubert, T. Lecroq, T. Merabti, S. J. Darmoni, and A. Rath "Mapping biomedical wordings utilizing normal language handling apparatuses and UMLS: Mapping the orphanet thesaurus to the MeSH," IRBM, vol. 31, pp. 221–225, Sep. 2010.

[3] A. M. Hao, S. Li, A. M. Hao, Q. Xia and Q. P. Zhao, "Deep learning for advanced calculation handling and investigation: An audit," J. Comput. Res. Create., vol. 56, no. 1, pp. 155–182, 2019.

[4] D. J. Berry, S. Fu, S. Sohn, C. C. Wyles, M. E. Tibbo, and H. Maradit Kremers, ''Use of natural language processing tools to identify and classify the chat transcript type and particular ID,'' J. Arthroplasty, vol. 34, no. 10, Oct. 2019.

[5] J. Yu, B. Zhang, Z. Kuang, D. Lin, W. Zhang and J. Fan, "Leveraging content affectability and client reliability to suggest fine-grained security settings for social picture sharing," IEEE Trans. Inf. Legal sciences Secured., vol. 13, no. 5, pp. 1317–1332, May 2018.

[6] William, Woods (1972) "The lunar sciences normal language information framework," BBN report, Bolt Beranek and Newman.

[7] K Javubar, Sathick, and Jaya, (2015) "Regular language to SQL age for semantic information extraction in friendly web sources," Indian Journal of Science and Technology, vol. 8, Issue 1, pp. 1–10.

[8] Iftikhar, Anum, Iftikhar, Erum, Mehmood and Muhammad Khalid, (2016) "Area explicit question age from regular language text," IEEE Sixth International Conference on Innovative Computing Technology (INTECH), pp. 502–506.