International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217698

Trends In Natural Language Processing : Scope And Challenges

Shreyashi Chowdhury, Asoke Nath
Computer Science, St. Xavier's College, Kolkata, West Bengal, India

ABSTRACT

Article Info
Volume 7, Issue 6
Page Number : 393-401

Publication Issue :
November-December-2021

Article History
Accepted : 15 Dec 2021
Published : 30 Dec 2021

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyse large amounts of natural language data. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them.NLP combines computational linguistics—rule-based modelling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment. Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation. This paper discusses on the various scope and challenges , current trends and future scopes of Natural Language Processing.

Keywords - Natural Language processing, Artificial Intelligence, Computer Linguistics, Speech Recognition, Natural Language Generation.

I. INTRODUCTION

As human beings started evolving and became social and organized animals, communication have played a huge role for mankind to come this far.With the advancement oftechnology, people recognized the importance of translation from one language to another and hoped to create a machine that could do this sort of translation automatically.

Natural Language processing (NLP) is the technology based on AI that enables the computers to understand human language whereas until some years earlier they were only capable of understanding mathematical language.[15].Some of important events in history of NLP:

1950 – NLP started with the publication of an article called "Machine and Intelligence" by Alan Turing.

1954- The Georgetown experiment involved fully automatic translation of over sixty sentences from Russian to English.

1960- The work of Chomsky and other linguists on language theory and generative syntax.

1990- Probabilistic and data driven models becomes standard.

2000- A large amount of spoken and textual data become available.[13]

That is, the evolution of NLP included the following major milestones: Symbolic NLP (1950s-Early 1990s), Statistical NLP (1990s–2010s), Neural NLP (Present)[6].NLP is divided into two parts, i.e. Natural Language Understanding and Natural Language Generation, to make processing easier.[13] Linguistics is the scientific study of language. It involvesanalyzing language form, language meaning and language in context [14], Its study includes- Sounds which refers to phonology, Word formation refers to morphology, Sentence structure refers to syntax, Meaning refers to semantics, Understanding refers to pragmatics [4]. Noah Chomsky, one of the first linguists of twelfth century that started syntactic theories, marked a unique position in the field of theoretical linguistics [9]. Some of the common tasks of NLP include Speech recognition,Part of speech tagging, Natural language generation[1]. But, there are some limitations in NLPwhich include Contextual words and phrases and homonyms, Ambiguity, Errors in text or speech[5].

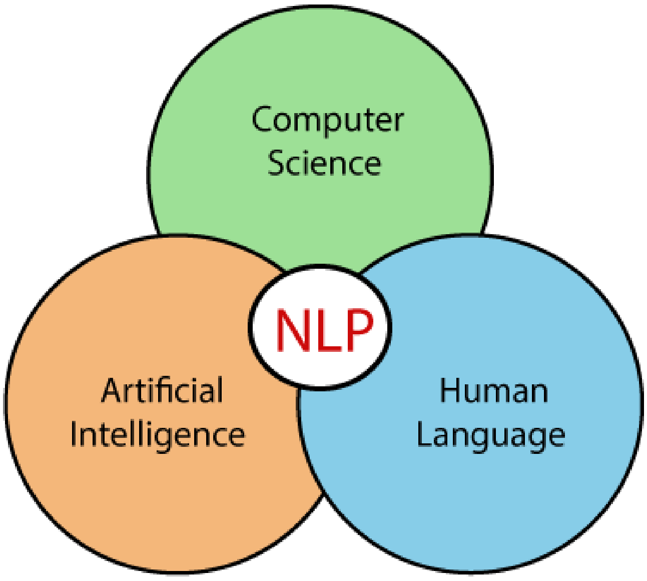


Fig.1 Natural Language Processing [17]

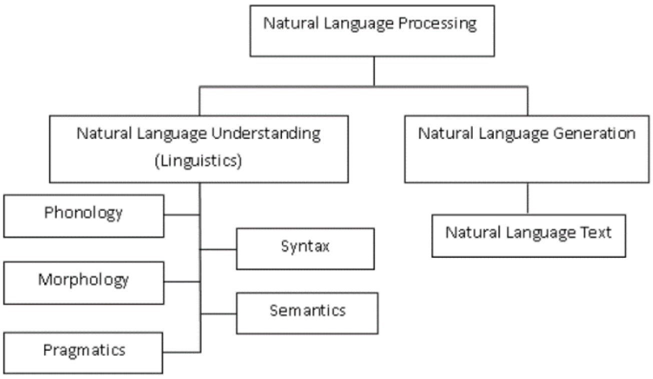

Natural Language Processing
Natural Language Understanding (Linguistics)
Natural Language Generation
Phonology
Morphology
Pragmatics
Syntax
Semantics
Natural Language Text

Fig 2. Classification of NLP [3]

II. NATURAL LANGUAGEPROCESSING

Natural Language Processing is a subfield of Artificial Intelligence and linguistics, devoted to make computers understand the statements or words written in human languages.

Natural Language Understanding:NLU is the process of conversion of unstructured data (data in human language) to structured data (form compatible for computer operations).[13] Its task is to understand and reason where input is a natural language.[4]

Natural Language Generation: NLG is the technology used by computers to communicate with humans in a way that they can understand.[13]It is a sub generation of natural

language processing. It is also referred to as text generation. [4]

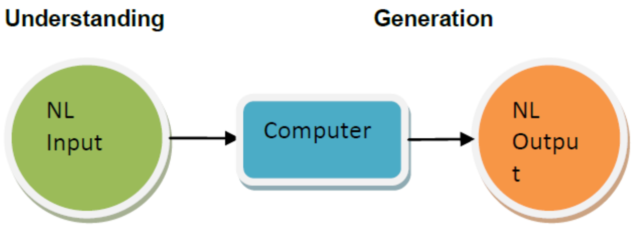


Fig 3. Natural Language Processing [4]

III. PHASES OF NATURAL LANGUAGE PROCESSING

There are 5 phases involved in natural language processing :

- **Morphological and Lexical Analysis**The lexicon of a language is its vocabulary that includes its words and expressions. Morphology depicts analysing, identifying and description of structure of words.[**4**]

  Lexical Analysis-The process of splitting a sentence into words or small units called "tokens" in order to identify the meaning of it and its relationship to the entire sentence.[**7**]
- **Syntactic Analysis**

  This involves analysation of the words in a sentence to depict the grammatical structure of the sentence. The words are transformed into structure that shows how the words are related to each other Eg. "the girl the go to the school". This would definitely be rejected by the English syntactic analyser.[**4**]
- **Semantic Analysis** This abstracts the dictionary meaning or the exact meaning from context. The structures which are created by the syntactic analyser are assigned meaning.Eg. "colourless blue idea".This would be rejected by the analyser as colourless blue do not make any sense altogether.[4]

  Output Transformation:Theprocess of generating an output based on the semantic analysis of the text or speech which fits the target of the application.[**7**]
- **Discourse Integration** The meaning of any single sentence depends upon the sentences that preceeds it and also invokes the meaning of the sentences that follow it .Eg the word "it" in the sentence "she wanted it" depends upon the prior discourse context. [**4**]
- **Pragmatic Analysis** It means abstracting or deriving the purposeful use of the language in situations importantly those aspects of language which require world knowledge the main focus is on what was said is reinterpreted on what it actuallymeans. Eg "closethewindow?" should have been interpreted as a request rather than order.[**12**]

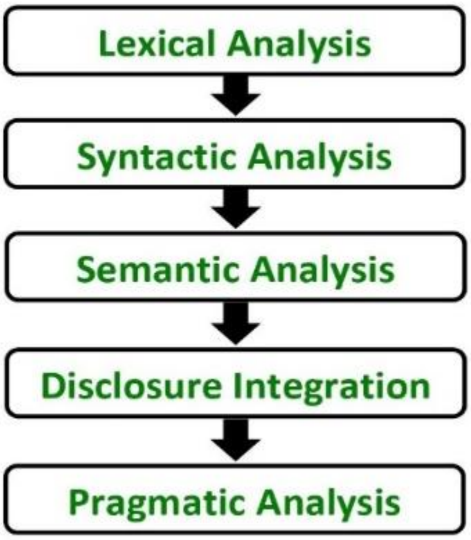


Fig **4**. Phases of NLP [**13**]

IV. APPROACH: NAÏVE BAYES CLASSIFIERS

Naïve Bayes Classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms that all share a common principle, that every feature being classified is independent of the value of any other feature.This algorithm is based on the conditional probability where, $P(A \mid B) = (P(A \cap B))/(P(B))$.

So for example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A Naive Bayes classifier considers each of these "features" (red, round, 3" in diameter) to contribute independently to the probability that the fruit is an apple, regardless of any correlations between features. Features, however, aren't always independent which is often seen as a shortcoming of the Naive Bayes algorithm and this is why it's labeled "naive". Although it's a relatively simple idea, Naive Bayes can often outperform other more sophisticated algorithms and is extremely useful in common applications like spam detection and document classification.[11]

Bayes' Theorem[16] Bayes' Theorem (also known as Bayes' rule) is a deceptively simple formula used to calculate conditional probability. The Theorem was named after English mathematician Thomas Bayes (1701-1761). The formal definition for the rule is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where:

P(A|B) – the probability of event A occurring, given event B has occurred

P(B|A) – the probability of event B occurring, given event A has occurred

P(A) – the probability of event A

P(B) – the probability of event B

Now, with regards to our dataset, we can apply Bayes' theorem in following way:

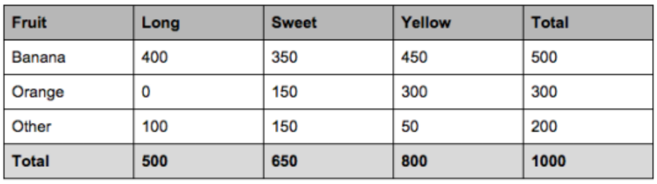
$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

where, y is class variable and X is a dependent feature vector (of size *n*) where:

$$X = (x_1, x_2, x_3, ..., x_n)$$

Explanation of Naïve Bayes Classification: [11]

Let's say we have data on 1000 pieces of fruit. The fruit being a Banana, Orange or some Other fruit and imagine we know 3 features of each fruit, whether it's long or not, sweet or not and yellow or not, as displayed in the table below:



Fruit	Long	Sweet	Yellow	Total
Banana	400	350	450	500
Orange	0	150	300	300
Other	100	150	50	200
Total	**500**	**650**	**800**	**1000**

Fig 5. Table of Example for Naïve Bayes Classifier

So from the table what do we already know?

- 50% of the fruits are bananas
- 30% are oranges
- 20% are other fruits

Based on our data set we can also say the following:

- From 500 bananas 400 (0.8) are Long, 350 (0.7) are Sweet and 450 (0.9) are Yellow
- Out of 300 oranges 0 are Long, 150 (0.5) are Sweet and 300 (1) are Yellow.
- From the remaining 200 fruits, 100 (0.5) are Long, 150 (0.75) are Sweet and 50 (0.25) are Yellow

Which should provide enough evidence to predict the class of another fruit as it's introduced.So let's say we're given the features of a piece of fruit and we need to predict the class. If we're told that the additional fruit is Long, Sweet and Yellow, we can classify it using the following formula and subbing in the values for each outcome, whether it's a Banana, an Orange or Other Fruit. The one with the highest probability (score) being the winner

Banana:

$$P(Banana|Long, Sweet, Yellow)$$
$$= \frac{P(Long|Banana) \cdot P(Sweet|Banana) \cdot P(Yellow|Banana) \cdot P(Banana)}{P(Long) \cdot P(Sweet) \cdot P(Yellow)}$$
$$= \frac{0.8 \times 0.7 \times 0.9 \times 0.5}{P(evidence)}$$
$$= \frac{0.252}{P(evidence)}$$

Orange:

$$P(Orange|Long, Sweet, Yellow) = 0$$

Other Fruit:

$$P(Other|Long, Sweet, Yellow)$$
$$= \frac{P(Long|Other) \cdot P(Sweet|Other) \cdot P(Yellow|Other) \cdot P(Other)}{P(Long) \cdot P(Sweet) \cdot P(Yellow)}$$
$$= \frac{0.5 \times 0.75 \times 0.25 \times 0.2}{P(evidence)}$$
$$= \frac{0.01875}{P(evidence)}$$

In this case, based on the higher score **0.01875 < 0.252** we can assume this Long, Sweet and Yellow fruit is, in fact, a Banana.

V. COMMON TASKS OF NLP

Several NLP tasks break down human text and voice data in ways that help the computer make sense of what it's ingesting. Some of these tasks include the following:

- **Speech recognition**, also called speech-to-text, is the task of reliably converting voice data into text data. Speech recognition is required for any application that follows voice commands or answers spoken questions. What makes speech recognition especially challenging is the way people talk—quickly, slurring words together, with varying emphasis and intonation, in different accents, and often using incorrect grammar. [1]
- **Part of speech tagging**, also called grammatical tagging, is the process of determining the part of speech of a particular word or piece of text based on its use and context. Part of speech identifies 'make' as a verb in 'I can make a paper plane,' and as a noun in 'What make of car do you own?'[1]
- **Word sense disambiguation** is the selection of the meaning of a word with multiplemeaningthrough a process of semantic analysis that determine the word that makes the most sense in the given context. For example, word sense disambiguation helps distinguish the meaning of the verb 'make' in 'make the grade' (achieve) vs. 'make a bet' (place). [1]
- **Named entity recognition,** or NEM, identifies words or phrases as useful entities. NEM identifies 'Kentucky' as a location or 'Fred' as a man's name.
- **Sentiment analysis** attempts to extract subjective qualities—attitudes, emotions, sarcasm, confusion, suspicion—from text.[1]

VI. USE CASES OF NLP

Natural language processing is the driving force behind machine intelligence in many modern real-world applications. Here are a few examples:

- **Spam detection:** The best spam detection technologies use NLP's text classification capabilities to scan emails for language that often indicates spam or phishing. These indicators can include overuse of financial terms, characteristic bad grammar, threatening language, inappropriate urgency, misspelled company names, and more.[1]

- **Machine translation:** Google Translate is an example of widely available NLP technology at work. Truly useful machine translation involves more than replacing words in one language with words of another. Effective translation has to capture accurately the meaning and tone of the input language and translate it to text with the same meaning and desired impact in the output language[1].Machine translation tools are making good progress in terms of accuracy.[10]
- **Virtual agents and chatbots:** Virtual agents such as Apple's Siri and Amazon's Alexa use speech recognition to recognize patterns in voice commands and natural language generation to respond with appropriate action or comments. Chatbots perform the same magic in response to typed text entries. The best of these also learn to recognize contextual clues about human requests and use them to provide even better responses or options over time. The next enhancement for these applications is question answering, the ability to respond to our questions—anticipated or not—with relevant and helpful answers in their own words.[1]

- **Sentiment Analysis in Social Media:**NLPhave become an essential business tool for uncovering hidden data insights from social media channels. Sentiment analysis can analyze language used in social media posts, responses, reviews, and more to extract attitudes and emotions in response to products, promotions, and events–information companies can use in product designs, advertising campaigns, and more.[1]
- **Fake News Detection:**The rapid rise in the popularity of social media platforms has not only fostered communication between various social groups but has also triggered the spread of fake news. NLP systems are frequently applied to detect fake information and provide statistics on its exposure. [1]

Fig 6. NLP Use Cases

VII. CURRENT TRENDS

Apart from Chatbots and Machine Translation, NLP has some more popular applications which are proving to be the game changer in today's time. Below is the description of some use cases which shows the power of NLP in the present era.

- **NLP in Health Care** Amazon Comprehend Medical services which are used to extract the disease conditions, can handle meditations sessions and can monitor the results of the treatment using clinical trial reports, electronic health records and using patient notes. This is an example of NLP in health analytics where using NLP the prediction of different diseases is possible using pattern recognition methods and patient 's speech and their electronic health record. [2]
- **Cognitive Analytics and NLP**Using NLP, the conversational frameworks are possible which can take commands by the medium of voice or by the medium of text. Using cognitive analytics, the automation of different technical processes are possible now such generation of a technical ticket related to a technical issue and also handling it in automated or semi-automated ways. The collaboration of these techniques can result in an automated process of handling technical issues inside an organization or providing the solution of some technical problems to the customer also in an automated manner. [2]
- **NLP in Recruitment:** NLP can also be used in both search and selection phases of Job Recruitment, in fact, the chatbot can also be used to handle the job-related query at Initial level which also includes identifying the required skills for a specific job and handling initial level tests and exams.[2]

VIII. LIMITATIONS AND CHALLENGES OF NLP

There are a number of challenges of natural language processing and most of them boil down to the fact that natural language is ever-evolving and always somewhat ambiguous. They include:

- **Errors in text and speech** Misspelled or misused words can create problems for text analysis. Autocorrect and grammar correction applications can handle common mistakes, but don't always understand the writer's intention. With spoken language, mispronunciations, different accents, stutters, etc., can be difficult for a machine to understand. [5]

- **Contextual words and phrases and homonyms.** The same words and phrases can have different meanings according the context of a sentence and many words – especially in English – have the exact same pronunciation but totally different meanings.
- **Homonyms** – two or more words that are pronounced the same but have different definitions – can be problematic for question answering and speech-to-text applications because they aren't written in text form. Usage of *their* and *there*, for example, is even a common problem for humans. [5]
- **Synonyms.** Synonyms can lead to issues similar to contextual understanding because we use many different words to express the same idea. Furthermore, some of these words may convey exactly the same meaning, while some may be levels of complexity (small, little, tiny, minute) and different people use synonyms to denote slightly different meanings within their personal vocabulary. So, for building NLP systems, it's important to include all of a word's possible meanings and all possible synonyms. [5]
- **Irony and sarcasm.**Irony and sarcasm present problems for machine learning models because they generally use words and phrases that, strictly by definition, may be positive or negative, but actually connote the opposite. [5]
- **Ambiguity.**Ambiguity in NLP refers to sentences and phrases that potentially have two or more possible interpretations.

a) **Lexical ambiguity:** a word that could be used as a verb, noun, or adjective.

b) **Semantic ambiguity:** the interpretation of a sentence in context. For example: *I saw the boy on the beach with my binoculars.* This could mean that I saw a boy through my binoculars or the boy had my binoculars with him.

c) **Syntactic ambiguity:** In the sentence above, this is what creates the confusion of meaning. The phrase *with my binoculars* could modify the verb, "saw," or the noun, "boy."[5]

IX. NLP PROJECTS

This section discusses the recent developmentsin the NLP projects implemented by various companies and these are as follows:

- **Eno A Natural Language Chatbot Launched by Capital One**Capital one announces chatbot for customers called Eno. Eno is a natural language chatbot that people socialize through texting. Capital one claims that Eno is First natural language SMS chatbot from a U.S. bank that allows customer to ask questions using natural language. Customers can interact with Eno asking questions about their savings and others using a text interface. Eno makes such an environment that it feels that a human is interacting. Eno provides a different platform than other brands that launch chatbots like Facebook Messenger and Skype.[3]
- **Meet the Pilot, world's first language translating earbuds** Waverly Labs' Pilot can already transliterate five spoken languages, English, French, Italian, Portuguese and Spanish, and seven written affixed languages, German, Hindi, Russian, Japanese, Arabic, Korean and Mandarin Chinese. The Pilot earpiece is connected via Bluetooth to the Pilot speech translation app, which uses speech recognition, machine translation and machine learning and speech synthesis technology. Simultaneously, the user will hear the translated version of the speech on the second earpiece. The earpieces can also be used for streaming music, answering voice calls and getting audio notifications.[3][8]

X. FUTURE SCOPE AND CONCLUSION

Natural language Processing is a growing technology, though it has its own challenges and limitations which mainly deals with data complexity, characteristics such as sparsity, diversityand the dynamic nature of the datasets. It still offers huge and wide-ranging benefits as well as vast scope and opportunities for engineers and industries to deal with many open challenges of implementing NLP systems. Today most of the AI systems use natural language processing such as Google Assistant, Netflix, Apple's Siri, and Grammarly which comes very handy.There is a belief that with the development in Natural Language Processing and Biometrics, machines like humanoid robots will acquire the capability to read the expressions of the faces as well as body languages and words alsofor example, a person is chatting with another person face to face. Humanoid robots are the necessity of this kind of communication as this can be the body to a programmed artificial soul. As the growth of NLP and Biometrics is gaining pace and accuracy as well, these technologies can give a whole new level to the research of Humanoid robots so that they can express themselves through movement, postures, and expressions.So from this we can say no matter whatever the limitations are there in the current times with new techniques and new technology cropping up every day, many of these barriers will be broken through in the coming years.

XI. REFERENCES

[1]. IBM Cloud Learn Hub, Natural Language Processing.OnlineRetrieved:October,2021.Available:https://www.ibm.com/cloud/learn/natural-language-processing

[2]. Xenostack Blog, Evolution ofNLP.(07 October2021). OnlineRetrieved:Novmber2021. Available:https://www.xenonstack.com/blog/evolution-of-nlp

[3]. Khurana, Diksha &Koli, Aditya &Khatter, Kiran & Singh, Sukhdev. (2017). Natural Language Processing: State of The Art, Current Trends and Challenges.

[4]. Abhimanyu Chopra, Abhinav Prashar, Chandresh Sain, (2013).Natural Language Processing.

[5]. Monkey Learn Blog, Limitations of NLP.(December 22nd, 2020)OnlineRetrieved: November2021.Available:https://monkeylearn.com/blog/natural-language-processing-challenges/

[6]. Dataversity, Brief history of NLP.(May 22, 2019)Online Retrieved:October,2021.Available:https://www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/

[7]. AI Multiple, Natural language processing. (August ,2021)Online Retrieved:October,2021.Available:https://research.aimultiple.com/nlp/

[8]. Meet the Pilot: Smart Earpiece Language Translator.(2016,May25).Online Retrieved: November 2021.Available: https://www.indiegogo.com/projects/meet-the-pilot-smart-earpiece-language-translator-headphones-travel

[9]. Chomsky, Noam. (1965). Aspects of the Theory of Syntax, Cambridge, Massachusetts: MIT Press.

[10]. Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., &Sawaf, H. (1997). Accelerated DP based search for statistical translation. In Eurospeech.

[11]. Data Science Central, Naïve Bayes Classification.(8th June 2015) Online Retrieved: November 2021 Available: https://www.datasciencecentral.com/profiles/blogs/naive-bayes-for-dummies-a-simple-explanation

[12]. "Natural Language Processing." Natural Language Processing RSS. N.p., n.d. (2017)

[13]. Pankaj Naruka, Rohit Yadav, Dr. Himanshu Arora, Monika Mehra.(2020) An Overview to Natural Language Processing.

[14]. A. GelbukhA. (2005) "Natural language processing", Fifth International Conference on Hybrid Intelligent Systems (HIS'05).

[15]. Krishna Prakash Kalyanathaya, D. Akila and P. Rajesh. (2019) "Advances in Natural Language Processing – A Survey of Current Research Trends, Development Tools and Industry Applications".

[16]. StatisticsHow To,Bayes Theorem.Online Retrieved:November,2021.Available:https://www.statisticshowto.com/probability-and-statistics/probability-main-index/bayes-theorem-problems/

[17]. Javapoint, History of Natural language processing ,OnlineRetrieved:October,2021 Available:https://www.javatpoint.com/nlp

Cite this article as :

Shreyashi Chowdhury, Asoke Nath, "Trends In Natural Language Processing : Scope And Challenges ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 6, pp. 393-401, November-December 2021. Available at
doi : https://doi.org/10.32628/CSEIT217698
Journal URL : https://ijsrcseit.com/CSEIT217698

AUTHOR PROFILE

Dr. Asoke Nath is working as Associate Professor in the Department of Computer Science, St. Xavier's College (Autonomous), Kolkata. He is engaged in research work in the field of Cryptography and Network Security, Steganography, Green Computing, Big data analytics, Li-Fi Technology, Mathematical modelling of Social Area Networks, MOOCs, Quantum Computing etc. He has published more than 257 research articles in different Journals and conference proceedings.

Shreyashi Chowdhury is a student of St. Xavier's College, currently pursuing M.Sc. in Computer Science. Her interests lie in the field of Natural Language Processing, Machine Learning, Human Computer Interaction, UI Design, Artificial Intelligence and real-world project implementation of these fields.