TAANI

Text Analytics Assistance by NIC

Edited by **MOHAN DAS VISWAM**

Text Analysis has become an essential part of information gathering from news, social media, and other documents to understand changing human sentiments, take corrective actions for effective policy implementation and intercept evolving dynamics of public engagement. It is founded on the bedrock of Natural Language Processing (NLP) algorithms, which helps machines to understand and process human languages. In this article, the concepts of Text Analytics and NLP along with their use case will be briefly discussed.

Sharmistha Dasgupta
Deputy Director General
& Head, CoE-AI
sharmi@nic.in

Rohit Kumar
Scientist-B
rohit.k12@nic.in

Usman
Scientist-B
usman.94@nic.in

Text Analysis has become an essential part of information gathering from news, social media, and other documents to understand changing human needs and take corrective actions for effective policy making. It is founded on the bedrock of Natural Language Processing, which helps machines to understand human languages. From predicting results with classifiers to generating texts that can fool humans, these tools have moved leaps and bounds. With regional languages in the picture, there are immense possibilities for these services other than simple transliteration and translation.

Technology Brief

NLP can be defined as the automatic manipulation of natural language, like speech and text by software. It helps machines in translation, search, and predictive text typing. NLP problems are as challenging as human language and filled with ambiguities which make it incredibly difficult to write software that can accurately determine the intended meaning of text or voice input. Several NLP tasks break down human text in a way that helps the computer make sense of what it is ingesting. Some of these tasks are:

Part of Speech (PoS) tagging: It is the process of determining the part of speech of a particular piece of text based on its use and context. For example, it identifies 'make' as a verb in 'I can make a paper plane,' and as a noun in 'What make of car do you own?'

Word Sense Disambiguation (WSD): WSD helps in the selection of correct meaning of a word from multiple definitions through a process of semantic analysis which determines word meaning that makes the most sense in the given context. For example, it can distinguish the difference in the word 'bank' in 'They bank on him to save the match' (depend - verb) compared to 'River bank is a beautiful place to do morning exercise' (place – noun).

Named Entity Recognition (NER): NER identifies words or phrases as useful entities. It helps to distinguish names of entities such as names of people, places, and things. It can be customised to identify other features of importance in text such as age, salaries and other data sets.

Sentiment Analysis (SA): SA attempts to extract subjective qualities such as attitudes, opinions, emotions, beliefs, and perspective from the text.

Features

- Context extraction using Attention mechanism
- Data interpretation by Language Modelling

- Customised Named Entity Extraction

Specifications

It is a known fact that machines do not understand text. They converse in the language of zeros and ones. For which, one needs to represent required information in encoded form. It can be done with the help of encoding tools such as Word2Vec and Bag of Words, which gives a probability to a particular set of words so that they can be represented in an array form and convey information such as frequency of their occurrence in the corpus. This is known as Term Frequency–Inverse Document Frequency (TF-IDF). These vectors are further used in text generation and machine translations. They are compatible with tensor maths which can be executed as matrix multiplications in tensor cores of a GPU in highly parallel fashion.

There is also a huge variety in document composition and textual context, including sources, format, language and grammar. Tackling this variety requires a range of methodologies:

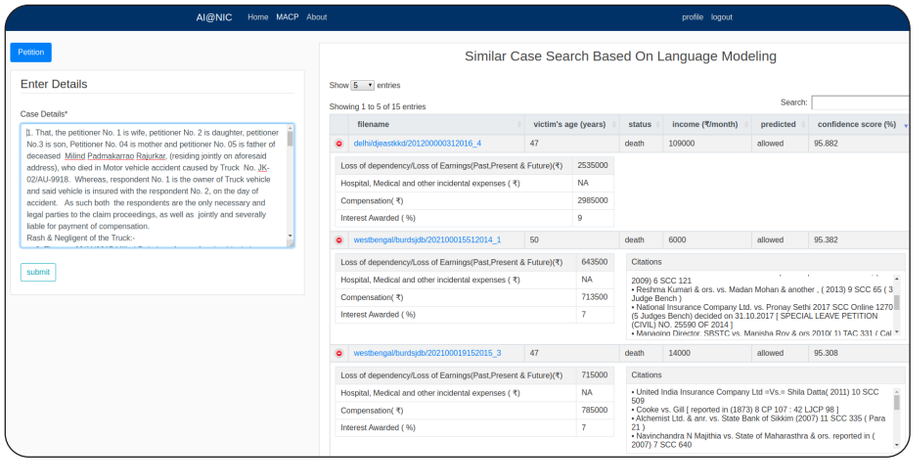
Text Pre-processing: Transformation of internal and external document formats (e.g., HTML, Word, PowerPoint, Excel, PDF text, PDF image) into a standardised searchable format, often require Optical Character Recognition (OCR) tool to extract texts from scanned pdfs. For the ability to process embedded tables within text, COE has developed an OCR based tool for text extraction from images and made them available to Judiciary MACP cases and CBSE.

Natural Language Toolkit (NLTK): Most of the time, we get unstructured, incorrectly formatted data, for which we may need to format it in order to make data ready for modelling purposes. NLTK is a python library developed by NIC CoE-AI that provides a number of tools for cleaning & making the data ready for modelling purposes.

Text Annotation: TA helps us to identify, tag and search in specific document sections which are important for training the AI model. For example, to use NER algorithms to detect credentials of a petitioner or respondent such as name, age, salary and dependents, one needs to annotate a number of cases to train the AI model. Several open source tools are available for doing text annotations.

Language Modelling: LM helps to determine the probability of a sequence of words in a text such as phrases or sentences, noun and verb groups together to form a relationship between themselves and make the computer capable of generating similar sentences.

Pattern recognition & Text Classification: To discover and identify categories of information, which are not easily defined with a dictionary approach like NER, Pattern Recognition helps to classify and categorise the text and help predict similarity scores in predictions. For text classification, we require tens of thousands of text annotations.



▲ Fig. 9.1: Motor accident claim petition

▲ Fig. 9.2: Text translation from 11 Indian language to English and vice versa

Model Training: For training a classifier model, a language model needs to be build, which is trained on a large corpus of data such as Wikipedia English in order to generate proper English sentences and teach domain specific lexicon for correct interpretation.

A language model attempts to learn the structure of natural language through hierarchical representations; thus, it contains both low-level features (word representations) and high-level features (semantic meaning). For example, After training a language model, it can form proper sentences and paragraphs like humans. It can be seen in the phrase "Petition was Signed", which is a totally fictitious piece.

Application Areas

In case of Motor Accident Claim Petition (MACP), over forty thousand case orders with representation from all High Courts were annotated in such a manner that the learning model is not biased. The cases were classified as accepted, partially accepted, partially rejected, rejected and settled. Also, citations and compensation were made available from case orders. Since, Hindi case orders also formed a sizable volume, the text translation was also carried out before annotations to give equal representation. (Refer Fig. 9.1)

In addition, NIC provides Matra Text Transliteration Services through Bharat API in 11 regional languages. Moreover, AI Panini - Text Translation based API Services Model has been deployed for translation from 11 Indian regional languages to English and vice versa. (Refer Fig. 9.2) These eleven languages are Hindi, Bengali, Gujarati, Marathi, Assamese, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil and Telugu. Similarly, AI Shruti provides Speech-to-Text and Text-to-Speech API Services in English and Hindi.

Benefits of Text analytics

- Allows interpretation of messages
- Topic modelling from text documents
- Text summarisation
- Q & A interpretation from text documents
- Text generation for predictive typing
- Sentiments analysis

Contact for more details

Rohit Kumar
CoE-AI, A3B8 Bay, NIC Hqrs.
CGO Complex, Lodhi Road, New Delhi - 110003
Email: rohit.k12@nic.in, Phone: 011-24305747