Vol 12 Issue 04 2023
ISSN NO: 2230-5807

Obtaining and Analyzing Data from Texts

Nadaf Farzana Peer Mohammed1, Fatima Rafique Shaikh2, Talawar Priya3,
Kamil Khan4, Sameena Jamadar5
1, 2, 3,4,5 Department of Computer Application, Abeda Inamdar Senior College, Pune, India

Abstract:
This paper provides an overview of the procedure for gathering and analysing data from texts. We are surrounded by textual information. You typically navigate through a lot of textual information just after you wake up in the form of text messages, emails, social media updates, and blog articles before you reach for your first cup of coffee. It might be tough to extract information from such vast amounts of text data. Large amounts of text data are handled by businesses from a variety of data sources, including apps, web pages, social media, customer reviews, support tickets, and call transcripts. Businesses use a technique called text mining to extract reliable, pertinent information from such massive amounts of text data. Text analysis software is used to carry out this process of information extraction from text data. Text mining tasks that are frequently performed include entity extraction, document summarization, text classification, text clustering, sentiment analysis, and the building of granular taxonomies.

Keywords*:* Text mining, Information extraction, Topic identification, Mobile learning.

I. INTRODUCTION:

This procedure entails a number of processes, including data collection, text pre processing, feature extraction, and analysis. Here is a rough outline of how to go about collecting and studying data from texts:

1. Data Acquisition:

- Determine where the text data that you wish to analyse came from. It might be anything that is text-based, including online pages, social media posts, academic articles, novels, and more.
- Depending on the source, you could need to use web scraping methods to collect the data, or you might have to use APIs or other data sources.

2. Text Pre processing:

- Text data must frequently be cleaned and normalised as part of pre processing before analysis.
- Typical pre processing procedures involve removing punctuation, changing text's case to lowercase, eliminating stop words (frequently used terms like "the," "is," etc.), and stemming or lemmatizing words to get rid of all but their most basic forms.

3. Feature Extraction:

- Text analysis requires the extraction of significant features.
- Term frequency-inverse document frequency (TF-IDF) scores and bag-of-words representation, where each document is represented as a vector of word frequencies, are common methodologies.
- Other methods include word embeddings, such as Word2Vec or GloVe, which encode words as dense vectors in a continuous space and capture semantic links.

4. Analysis:

- Depending on your goals, you can undertake a variety of analyses after obtaining, pre processing, and extracting the data's pertinent aspects.
- Examples include named entity recognition, sentiment analysis, topic modelling, text categorization, and information extraction.

- Models can be trained on labelled data using machine learning or deep learning algorithms for supervised tasks like clustering or dimensionality reduction, or unsupervised methods.

5. **Interpretation and Visualization:**

- In order to get insights, it is crucial to interpret and visualise the outcomes of the data analysis.
- Data visualisation methods that can be used to understand patterns, trends, and relationships in the data include word clouds, bar charts, heat maps, and network graphs.

6. **Iterative Process:**

- The process of collecting and studying data from texts is frequently iterative. Based on preliminary findings and new information learned, you might need to modify your pre processing procedures, feature extraction strategies, or analysis methodologies.

II. TEXT MINING TECHNIQUES:

1. **Information Extraction (IE)**

- It is a method for obtaining important data from vast amounts of data. IE involves tasks including tokenization, named entity identification, sentence segmentation, and part-of-speech assignments and is the first stage for systems to understand unstructured text by identifying key phrases and relationships within text.
- IE systems are used to extract specific data, properties, and entities from documents and identify their relationships. Following this, the collected corpora are gathered into related databases for further processing. Precision and recall processes are used to the extracted data in order to inspect and assess the relevant data/outcomes.
- To process information extraction techniques for maximising results, in-depth and wide knowledge of the related field is required.

2. **Information Retrieval (IR)**

- The practise of collecting important information and associated patterns from a given set of words or phrases is known as information retrieval (IR). Different algorithms are used in information retrieval to monitor user behaviour and find pertinent data and information as a result.
- For instance, Google Search Engine continually employs information retrieval techniques to find relevant documents based on terms entered into the search bar. Search engines use query-based algorithms for this reason in order to retain trends and produce more relevant results. Following that, search engines offer users more pertinent and reliable information based on their search requirements.

3. **NATURAL LANGUAGE PROCESSING**

- NLP is concerned with the automatic processing and analysis of unstructured textual information. It enables computers to read by examining the syntax and sentence structure. As seen below, it does several types of analysis including NER, summarization, and sentiment analysis.

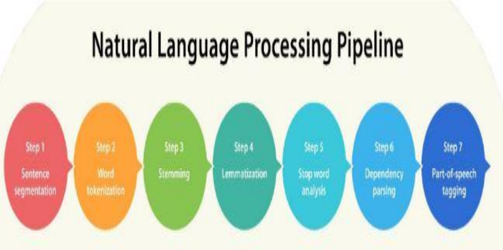


Fig 1: NLP Pipeline

- **Summarization:** to provide a summary of a large amount of content in order to create a clear, succinct, and understandable overview of a document's main points.

- **Text categorization:** to categorise synonyms and abbreviations when categorising text documents based on analysis and predetermined topics or groups. Another name for it is text classification.
- **Sentiment analysis:** to identify favourable or unfavourable sentiment from internal and external data sources and enable users to track changes in client behaviour over a predetermined time frame. Sentiment analysis is used to gather pertinent data on consumer views of brands, goods, and services, which encourages businesses to engage with their target market to enhance workflows, the user experience, and customer satisfaction.

4. CLUSTERING

- The clustering method is an unsupervised procedure that uses different clustering algorithms to categorise text texts into groups. Similar phrases or patterns are organised and extracted from various texts through the process of clustering, which is done both top-down and bottom-up.
- As a result, different partitions, known as clusters, are created; each cluster contains a certain number of documents. Each document in a cluster has fairly similar content; however the content of documents in various clusters differs, improving the quality of clustering.
- Each document's subjects are tracked by a basic clustering algorithm, which also assigns weights based on how well the documents fit into each cluster.
- A good clustering approach produces excellent clusters with high intra-cluster similarity and low inter-cluster similarity. The quality of a clustering result depends on the similarity measures of text content employed by the clustering method and its implementation.
- Clustering is distinct from categorization since text contents are grouped together without prior knowledge of the classifications in clustering. The fundamental benefit of text content clustering is that it can be pertinent to numerous classes.
- For analysing unstructured text documents, various clustering algorithms include hierarchical, distribution, density centroid, and k-means clustering.

5. CATEGORIZATION

- The categorization approach assigns independent (free format) text documents to one or more categories. Categorization is regarded as a supervised learning method because it depends on input-output examples to distinguish new texts. Each text document is given a predefined class based on the substance of the text.
- Text is categorised using techniques including pre-processing, indexing, dimensionality reduction, and classification with the goal of training classifiers based on recognised examples, after which unrecognised examples would be automatically categorised. The great dimensionality of the feature space presents another challenge for text categorization.
- The closest neighbour classifier, decision trees, support vector machines, and naive Bayesian classifier are some helpful analytical classification methods that can be used to categorise text. Document organisation, spam filtering, SMS classification, and hierarchical web page categorization are all applications that fall within the category of categorization.

6. VISUALIZATION

- Visualisation techniques can enhance and clarify the analysis of pertinent data. Text flags are used to indicate a document's category and colours to indicate the density of documents in order to outline certain papers or collections of documents.

- This approach arranges substantial text sources in a visual hierarchy to enable user interaction with the documents through scaling and diving. For instance, the government employs information visualisation to find criminal information and uncover terrorist networks.

Three steps make up the visualisation technique procedure;

1. **Data preparation:** In this step, the original data for the visualisation are identified, acquired, and an original data space is created.
2. **Data Analysis and Extraction:** Data analysis and extraction refers to the process of analysing and extracting visualisation data from the original data in order to create a visualisation data space.
3. **Visualization Mapping:** A few mapping techniques are used in this stage to translate the visualisation data space to the visualisation target.

7. TEXT SUMMARIZATION

Text summarization helps determine whether a long document satisfies the user's needs and whether it is worthwhile to read for additional information. As a result, text summarization may be replaced by groups of documents with the fundamental goal of reducing the length, details, and complexity of a document while maintaining significant points and actual meaning.

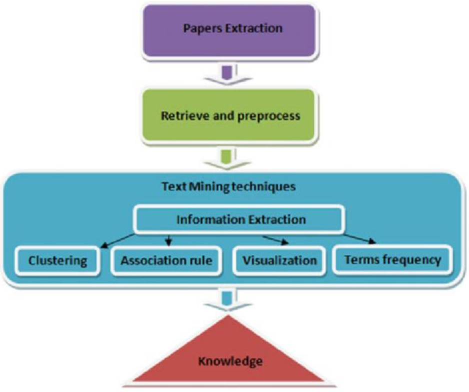
Text summary software manages and summarises a lengthy text document faster than people do whenever a user reads the first paragraph. It can be divided into two categories;

- **Abstractive Summarization:** It produces a distinct view of the text's core topics and illustrates them using everyday language. It uses linguistics techniques to interpret, rewrite, and explain text in precise form.
- **Extractive Summarization:** In order to identify the sentences that need to be extracted, these are carried out by deriving main text segments and depending on statistical analysis of text properties such word/phrase frequency, position, or suggested terms.

Tex summarization, in particular, is a three-step procedure;

- **Pre-processing:** The actual text is represented structurally in this stage. Some techniques used for pre-processing include tokenization, stop word elimination, and stemming.
- **Processing:** In order to convert and comprehend summary structure from text structure, algorithms are used.
- **Development state:** In this stage, the summary structure's final summary is retrieved.

Fig.2 Text mining processing framework



CONCLUSION

Effective methods must be used to analyse the data and extract pertinent information from it due to the growing amount of text data. We know that several text mining approaches are employed to effectively extract the pertinent information from a variety of textual data sources and are continuously used to enhance the text mining process.

To make the text mining process simple and effective, the right techniques and tools should be chosen and used in accordance with the business challenges and requirements.

References :

[1] https://www.researchgate.net/publication/321150349_Using_Text_Mining_Techniques_for_Extracting_Information_from_Research_Articles

[2] https://www.researchgate.net/publication/313075362_A_Survey_of_Text_Mining_in_Social_Media_Facebook_and_Twitter_Perspectives

[3] https://www.ijcaonline.org/archives/volume85/number17/14937-3507

[4] https://www.researchgate.net/publication/283668827_An_Approach_for_Sentiment_Analysis_on_Social_Networking_Sites

[5] http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

[6] https://www.temida.si/~bojan/MPS/materials/Data_preparation_for_data_mining.pdf

[7] https://www.scirp.org/(S(vtj3fa45qm1ean45vvffcz55))/reference/ReferencesPapers.aspx?ReferenceID=1863384

[8] https://www.mdpi.com/2071-1050/13/2/917

[9] https://www.researchgate.net/profile/Sameer-Mohammad-4/publication/354178057_A_Survey_on_Text_Mining_and_Sentiment_Analysis_for_Unstructured_Web_Data/links/612a05fbc69a4e4879605578/A-Survey-on-Text-Mining-and-Sentiment-Analysis-for-Unstructured-Web-Data.pdf

[10] https://www.sciencedirect.com/science/article/abs/pii/S1566253516302354

[11] https://www.semanticscholar.org/paper/A-review-on-text-mining-Zhang-Chen/d46017f8dc834a3fda4c4b1c2f5ffa3db95a9d2e

[12] https://www.turing.com/kb/natural-language-processing-function-in-ai