IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

"A LITERATURE PAPER ON TEXT SUMMARIZATION AND VISUALIZATION"

Ms. Preeti D. Chandanshive
Ms. Sneha D. Shinde
Ms. Mrunali V. Gaikwad
sMs. Vaishnavi H. Biradar
&
Prof. Priti . Chorde
Dept of Computer Engineering.
Vidya Prasarini Sabha's College of Enhineering andTechnology Lonavala , Maharshtra India

Abstract :

Text Analysis is the discovery and extraction of interesting, non-trivial knowledge from free or unstructured text. This encompasses everything from information retrieval (i.e., document or web site retrieval) to text classification and clustering, to (somewhat more recently) entity relation, and event extraction. Natural language processing (NLP), is the attempt to extract a fuller meaning representation from free text. This can be put roughly as figuring out who did what to whom, when, where, how and why. This paper presents a case study of implementing computational methods like Natural Language Processing (NLP) to perform Text Analytics and Visualization on given data. This data is sometimes available in unstructured textual format and thus they are a part of big data requiring analytics to derive insights from it. In this experiment, a significantly large volume of Text is analyzed and graphical visualizations are generated such as Word cloud, Mendenhall Curve, Tokenization, Graph, Processed Text, and Name Entity Recognition (NER) using various Python Libraries. Therefore, this paper is an attempt to provide detailed assessment about data, also to make proper

Keywords : Text Filtering, Extraction of words and shorten the results

INTRODUCTION

1.1 MOTIVATION

1. Understanding Customer Feedback: One of the key motivations for text analysis is to analyze customer feedback. Companies can use NLP to analyze customer reviews, feedback, and comments to gain insights into customer sentiment and preferences.

2. Identifying Key Topics: Another motivation is to identify key topics within large volumes of text. This can be useful for organizations that need to quickly identify key issues and concerns within large volumes of text, such as social media feeds, news articles, or research papers.

3. Improving Search: NLP can also be used to improve search results. By analyzing the text within documents, NLP algorithms can help identify related documents, key topics, and relevant keywords, making search results more accurate and relevant.

4. Automating Text Analysis: NLP can also be used to automate text analysis, allowing organizations to analyze large volumes of text quickly and efficiently. This can be especially useful for companies that need to analyze large amounts of data on a regular basis, such as financial institutions or healthcare providers.

5. Visualizing Data: Finally, text analysis can be used to create visualizations that help users understand and interpret large

volumes of data. By creating charts, graphs, and other visualizations, NLP can help users quickly identify key trends and insights within complex

dataset

OBJECTIVES

1. Understanding the content: Text analysis can help extract meaning and identify key themes or topics in a large body of text. Text visualization can then help make these themes and topics more visible and easily understandable.

2. Identifying patterns and trends: Text analysis and visualization can be used to identify patterns and trends over time, such as changes in sentiment towards a particular topic or changes in the language used to describe certain issues.

3. Improving decision-making: By extracting insights from text data, text analysis and

visualization can help decision-makers make more informed and data-driven decisions.

4. Enhancing communication: Text visualization can be used to communicate complex information in a more accessible and intuitive format, making it easier for people to understand and engage with the information being presented.

TEXT SRCUTINY TECHNIQUES:

1. Text Summarization - Extractive Summarization: Selects important sentences or phrases from the original text to create a summary.- Abstractive Summarization: Generates summaries by paraphrasing and rephrasing the content using natural language generation techniques.

Sentiment Analysis:- Determines the sentiment (positive, negative, neutral expressed in a piece of text. - Can be used for understanding public opinion, customer feedback analysis, and brand sentiment monitoring.

1. Named Entity Recognition (NER): - Identifies and classifies entities such as names of people, organizations, locations, dates, and more in text - Crucial for extracting structured information from unstructured text.

2. Topic Modeling:- Divides a collection of documents into topics based on the underlying thematic content.- Common methods include Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF).

3. Text Classification:- Assigns predefined categories or labels to text documents based on their content.- Used for spam detection, sentiment classification, and categorizing news articles, among others.

4. Dependency Parsing: - Analyzes the grammatical structure of a sentence by identifying relationships between words.

5. Word Embeddings:- Represents words as dense vectors in a high-dimensional space, capturing semantic relationships between words.- Techniques like Word2Vec, GloVe, and FastText are commonly used for creating word embeddig

TEXT VISUALIZATION TECHNIQUES:

Text visualization is one of the most important tools for text mining due to its readability to both human and machines.

Text visualization is mainly achieved through the use of graph, chart, word cloud, map, network, timeline, etc. It is these visualized results that make it possible for humans to read the most important aspects of a huge amount of information

SCOPE OF WORK

- Collecting the relevant text data to be analyzed and visualized, which may involve web scraping, data mining, or other methods.
- Cleaning and formatting the text data to ensure that it is ready for analysis, which may involve
removing stop words, stemming or lemmatization, and removing punctuation or special
characters.
- Applying various text analysis techniques to extract meaningful insights from the text data, such as sentiment analysis, topic modeling, and entity recognition.
- Creating visual representations of the text data to help communicate insights and trends to stakeholders, such as word clouds, bar charts, heatmaps, and network diagrams.
- Assessing the effectiveness of the text analysis and visualization techniques used, and

refining
the approach as needed to ensure that it meets the needs of the organization and its stakeholders.

• Deploying the text analysis and visualization tools within the organization, and providing
training and support to ensure that stakeholders can effectively use and interpret the data.

• Ongoing monitoring and maintenance of the text analysis and visualization tools to ensure that

they continue to provide accurate and useful insights over

− To detect the given text as input, perform analysis on the data and show the score of the polarity of input text.
− The input will be taken from the user in string format. After inputting the string,
the approach used in this project classify/tokenize the text in tokens. When
tokenization is completed, it starts operation of tagging to each token and then
evaluate it.

LIMITATIONS:

• Contextual limitations: Text analysis and visualization tools often rely on statistical
and computational methods that may not be able to fully capture the nuances of
language, tone, and cultural context.

• Bias: Text analysis and visualization tools can be biased if the data they are trained on
is biased. For example, if a tool is trained on a dataset that has a particular racial or
gender bias, the tool may produce biased results when applied to other datasets.

• Data quality: The accuracy and completeness of the data used in text analysis and
visualization can significantly impact the results. If the data is incomplete or contains errors, the analysis and visualization may be inaccurate or misleading.

• Lack of human input: Text analysis and visualization tools can automate much of the
analysis process, but they lack the nuanced understanding that humans can bring to the
data. This can result in a loss of insight and understanding that can only be gained through human interpretation

APPLICATIONS:

Dialogue Generation: This category focuses on generating text in the form of a conversation between two or more agents. Dialogue generation systems are used in various applications, such as chatbots, virtual assistants, and conversational AI systems. These systems use dialogue history, user input, and context to generate appropriate and coherent responses. P2-BOT is a transmitter–receiver-based framework that aims to explicitly model understanding in chat dialogue systems through mutual persona perception, resulting in improved personalized dialogue generation based on both automated metrics and human evaluation.

Code Generation: This category focuses on generating code based on a given input, such as a natural language description of a software problem. Code generation systems are used in software development to automate repetitive tasks, improve productivity, and reduce errors. These systems can be trained to use expert knowledge, and can be specialized for a single programming language, such as SQL, or trained on a large corpus to support various programming languages and different programming paradigms.

Data-to-Text Generation: This category focuses on generating natural language text from structured data such as tables, databases, or graphs. Data-to-text generation systems are used in various applications, such as news reporting, data visualization, and technical writing.

These systems use natural language generation techniques to convert data into human-readable text, taking into account the context, target audience, and purpose of the text.

Control Prefixes extend prefix tuning by incorporating input-dependent information into a pre-trained transformer through attribute-level

Query-Focused Summarization: This subtask focuses on summarizing a document based on a specific query or topic. Query-focused

summarization methods typically use information retrieval techniques to identify the most relevant sentences or phrases in a document and present them as a summary. Baumel et al. [35] introduced a pre-inference step involving computing the relevance between the query and each sentence of the document. The quality of summarization has been shown to improve when incorporating this form of relevance as an additional input.

TOOLS AND FRAMEWORK:

1. TextBlob

TextBlob is a Python library that is used to process textual data, with a primary focus on making common text- processing functions accessible via easy-to-use interfaces.

Objects within TextBlob can be used as Python strings that can deliver NLP functionality to help build text analysis applications.

2. SpaCy

This open-source Python NLP library has established itself as the go-to library for production usage, simplifying the development of applications that focus on processing significant volumes of text in a short space of time.

3. Natural Language Toolkit (NLTK)

NLTK consists of a wide range of text-processing libraries and is one of the most popular Python platforms for processing human language data and text analysis. Favoured by experienced NLP developers and beginners, this toolkit provides a simple introduction to programming applications that are designed for language processing purposes.

4. Genism

Genism is a bespoke Python library that has been designed to deliver document indexing, topic modelling and retrieval solutions, using a large number of Corpora resources.

ALGORITHMS

1. Naive bayes (NB): Naive Bayes is a classification algorithm based on Bayes' theorem. It is a simple yet powerful algorithm that is commonly used in NLP and text classification tasks.

The basic idea behind Naive Bayes is to calculate the probability of a document belonging to a particular class based on the occurrence of certain features in the document.

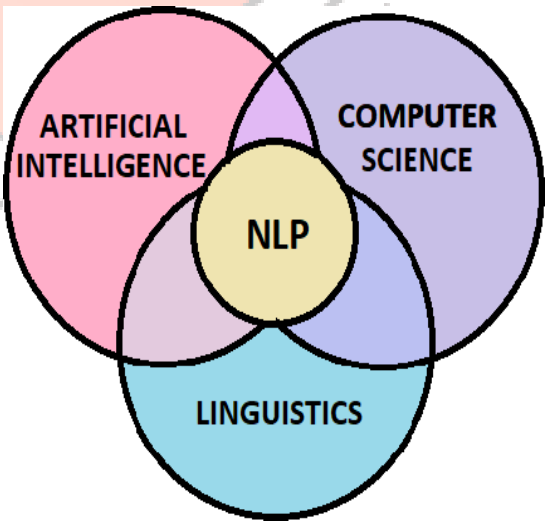
Naive Bayes is a relatively fast and simple algorithm that classification tasks where the number of features is large compared to the number of training examples.

However, it can suffer from the "zero- frequency" problem if a feature appears in the test data but not in the training data, which can cause the conditional probability to be zero. There are several variants of the Naive Bayes algorithm, such as Laplace smoothing and Bernoulli Naive Bayes, that can help to mitigate this problem.

2. Support vector machine (svm):It is an algorithm that can divide a vector space of tagged texts into two subspaces: one space that contains most of the vectors that belong to a given tag and another subspace that contains most of the vectors that do not belong to that one tag.

SVM (Support Vector Machine) is a popular machine learning algorithm used for classification and regression tasks.

It is a supervised learning algorithm that works by finding the best hyperplane that separates the data into different classes.



FUTURE ENHANCEMENT:

We can expect to see more advanced techniques for understanding and interpreting human language.

This could include more accurate sentiment analysis, entity recognition, and semantic analysis. By Integrating machine learning and AI we could include more accurate predictive analytics, personalized recommendations, and automated content creation.

Text analysis and visualization tools can automate much of the analysis process, but they lack the nuanced understanding that humans can bring to the data. This can result in a loss of insight and understanding that can only be gained through human interpretation.

CONCULSION:

Text analysis and visualization have the potential to revolutionize the way organizations approach data analysis and decision-making, and can provide a significant competitive advantage for those who are able to harness their power effectively.

In conclusion, this survey paper has provided a comprehensive overview of the dynamic landscape of text scrutiny and visualization. Through an exploration of various techniques and methodologies, we have unveiled the pivotal role that these approaches play in unraveling the hidden insights concealed within vast oceans of textual data.

Text scrutiny techniques have evolved to decipher the intricate tapestry of language, enabling us to extract meaning, sentiment, entities, and topics from unstructured text. The spectrum of methods, ranging from sentiment analysis to named entity recognition, empowers us to decode the semantics and syntax that underlie human communication. By engaging in an in-depth analysis of these techniques, we have underscored their

REFERENCES

1. Diksha Khurana, Aditya Koli, Kiran khatter & Sukhdev Singh " Natural language
2. processing: state of the art, current trends and challenges", 10.1007/s11042-022- 13428-4
3. John Risch, Shawn J bohn, Anne kao, Steve Poteet " Text Visualization",DOI
4. Sarthak J Shetty , Vijay Ramesh, " An open- source Python package for scientifictext analysis" , https://doi.org/10.1002/ece3.8098
5. Atharva Deshpande, Vijay Shetty , An open python package