IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A SURVEY ON TECHNIQUES, METHODS, AND APPLICATIONS OF TEXT ANALYSIS

[1]Aanal S Raval, [2]Agrima Lohia
[1,2]Academic Associates
Area-Information Systems
IIM, Ahmedabad, India

Abstract: Text analysis is a popular area of research for the retrieval of essential information for further surveys. The meteoric advancement of digital text data has led to high volumes with diversity. There has always been an issue for the selection of proper techniques to retrieve underlying patterns and trends from unstructured and vast data. Different techniques for this textual data analysis like preprocessing, natural language processing, sentiment analysis, and classification are discussed here for proper selection according to the decision making.

Key terms: *text preprocessing, NLP, sentiment analysis, statistical methods, text classification, regression, topic modelling.*

I. INTRODUCTION:

Text analysis is the process of detecting underlying patterns and trends, as well as extracting worthy information out of unstructured and vast data. It can be anything like analyzing text written by customers in a customer survey, with the focus on finding common themes and trends, examining the customer feedback to inform the business on taking strategic action, to improve customer experience, etc [19]. To make this analysis more efficient, NLP and preprocessing steps are inevitable to find interest in an enormous amount of text [19].

II. OVERVIEW OF PHASE SEGMENTS:

The first phase is always of low-level text preprocessing followed by some NLP steps. General problems encountered in this phase are missing data, Manual input, Data inconsistency, Regional formats, Numerical units, Wrong data types, File manipulation, Missing anonymization [24]. NLP involves problems like ambiguity, errors in text, domain-specific and low resource language, contextual words, and synonyms.[25] After NLP, if required statistical analysis can be applied at this stage. Then topic modeling is followed if needed. Thereafter sentiment analysis and/or classification can be performed. Training and Sampling should be concerned with classification.

III. METHODOLOGY

3.1 Text Preprocessing and NLP techniques:

Its a task of transferring any text to a format that is foreseeable by removing noisy parts for a convenient survey that includes primarily processes like stop word removal, elimination of null values, lowercasing, and uppercasing, removal of punctuation marks, URL, HTML tags, removing numerics, removal of rare words or frequent words or whitespaces, spelling correction and conversion to ASCII characters.

Any among these or a combination from above is generally preferred for basic preprocessing. Sometimes for a more desirable format, NLP steps are applied. Some of them are summarized here.

Nlp step is commonly applied for extraction of information for further computations in algorithms, specifically for dealing with meaningful portions, syntactic as well as semantic processing, and dealing with segmentation. It can include tokenization (retrieving smaller base units called tokens), normalization (conversion of a given list of text to a uniform sequence), or even clustering methods like K-means or hierarchical methods for grouping, according to need. Commonly used methods are explored hereby.

3.1.1 Stemming: It is the reduction of a word to its stem which affixes to suffixes and prefixes. Its common application is seen in queries and search engines.[1]

3.1.2 Lemmatization: Irrespective like in stemming that focused on the removal of some characters for root word, this method also keeps the context and extracts a meaningful base word eliminating problems of incorrect meaning and spelling, thus providing high-quality information from text.

3.1.3 Word embeddings: [2]These are a type of word representation that allows words with similar meaning to have a similar representation, where individual one hot encoded-words are represented as real-valued vectors.

3.1.3.1 Word2vec: Its commonly known types are: CBOW (for predicting a particular base word) and Skip-gram (for predicting surrounding words for a given base word).

3.1.3.2 GloVe: It is a combination of local context-based learning and LSA (Latent semantic analysis)

3.1.3.3 IWE: [3][4] Intelligent Word embedding combines word2vec with semantic dictionary mapping to tackle problems with out of vocabulary words and morphologically similar words.

3.1.3.4 Contextualized word representation: it models both syntax, semantics of word, and how these uses vary across linguistic contexts. So the representation for each word depends on the entire context in which it is used.[12]

3.1.3.5 FastText: It is a library for efficient learning of word representations and sentence classification.[13]

3.1.3.6 Document embedding: the corresponding algorithms used here are “distributed memory” (DM) and “distributed bag of words” (DBOW). DBOW is similar to skip-gram that obtains paragraph vectors by training a neural network on the prediction of the probability of words in a paragraph given a randomly sampled word from the paragraph. DM is a memory that finds what is missing from the current context. So doc2vec intends to represent the concept of document[14].

Sometimes, prior to embeddings, CountVectorizer (or one-hot encoding) is applied to convert text to token counts, if numeric representation is needed.

3.1.4 POS tagging:[5] part-of-speech tagging is the technique to assign labels to each token in text, to indicate its part of speech and even other grammatical connotations.

3.1.5 Named Entity Recognition and Disambiguation: [6]Objective of NER is to locate and classify named entities in text into predefined categories such as the names of persons, organizations, locations, events, expressions of times, quantities, monetary values, percentages, etc. And identifying the correct meaning of mentions in a sentence can be done by NED[7].

3.1.6 Semantic text similarity: Its analysis of similarity w.r.t. meaning and essence of the text rather than only syntax[7], which can be string based (character and term-based), Corpus-based, or knowledge-based.[8]

3.1.7 Language Identification: Concerns with giving labels to text[9]. The identification methods explored as per[9] can be seen as a special case for text classification.

3.1.8 Text Summarization:[7] Task here is to identify the important points without altering the meaning from the given text and making it straightforward.

3.1.9 Data Augmentation: Sometimes it happens that acquiring and labeling additional observations can be an expensive and time-consuming process, So to make datasets larger, Data augmentation techniques are used to generate additional, synthetic data using the data you have[18].

3.2 Sentiment Analysis:

It can be said as recognizing the state of given information that includes text analysis, biometrics, natural language processing (NLP), and computational linguistics which in turn can be applied to social media monitoring, market research, customer attitudes, and preference knowledge and knowing public opinion on specific topics on large scale[10]. Standard sentiment analysis simply states two or three class classifications as positive and negative or can include neutral.

Based on a wide range of emotions and expressions, here are explored some important types of sentiment analysis[10]:

3.2.1 Fine-Grained: This type of sentiment analysis gives polarity precision to normal categories like very positive, positive, neutral, negative, or very negative and so on, as well as can provide scale rating.

3.2.2 Aspect Based: This type allows retrieval of sentiments regarding the aspects of a given sentence, for instance, to determine that a customer has posted something "positive" on a particular brand name.

3.2.3 Emotion Detection: This type allows to detection of emotions like anger, sadness, happiness, frustration, fear, worry, panic, etc. using lexicons.

3.2.4 Intent Detection: It relates text or expression with a particular intent, For example, words buy or acquire are related to purchase.[11] For more clear understanding, consider an example in [11] "I want to book a flight from New York to Las Vegas, but my card has been declined" so an intent classifier would categorize it as intent to Book a Flight, and a text extractor would extract the entities New York and Las Vegas.

3.3 Statistical methods:

These methods include word frequency count, collation, concordance and TF-IDF. Word frequency identifies the most frequently used words or expressions in a specific piece of text, which can address problems like identifying success are.[16]. Talking about concordance, it has input information about the total number of documents present, matching of several number of documents containing the queried word, and number and type of tokens. So it finds the queried word in a text and displays the context in which this word is used. Results in a single color come from the same document. The widget can output selected documents for further analysis or a table of concordances for the queried word[15]. For example, "your license has been issued" refers to "it is supplied", "There is an issue with my laptop." refers to "problem"[16]. Now Collation is a technique that allows identifying words that occur together. For its instance, bigrams and trigrams can be considered[16]. Term frequency refers to the number of particular words that appear in a given document. Inverse document frequency refers to the frequency of that word across the total number of given documents. So TF-IDF is a measure that evaluates how relevant a word is to a document in a collection of documents and is obtained by multiplication of the above two given metrics[17].

3.4 Text Analysis (Regression, Classification, and Extraction):

These tasks include keyword spotting, manual rules, categorization, Topic modelling, assigning labels, assignment of probabilities (for example, what is the probability that a given users' social media account is interested in 'politics') as well as extraction of entities like company or brand name, people, topics, etc.

3.4.1 Word Spotting[19]: This is capable of doing tasks like handwriting recognition i.e., spotting which word a person, a doctor perhaps, has written.

3.4.2 Categorization: For tasks like classification to different categories either for political, market, product survey, or knowing customers' interest or classification of something to related topics, this approach is generally preferred, using different algorithms like naive bayes, linear classifiers, SVM, bagging-boosting models (Ensemble learning and NN.

3.4.3 Regression: [20] regression models have many applications, particularly in financial forecasting, trend analysis, marketing, time series prediction and even drug response modeling.

3.4.4 Topic modelling: It doesn't need any input other than raw text and it learns by observing which words appear alongside other words, thus capturing information with help of probability statistics. [19] The model used here is LDA (Latent Dirichlet Allocation.).

REFERENCES:

[1]. *TechTarget. Stemming* [online] Available at:

What is stemming? - Definition from WhatIs.com.

[2]. *Machine Learning Mastery*. 2017. *What are Word Embeddings for Text?* [online] Available at:

87 Responses to What Are Word Embeddings for Text?

[3] *"Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort",*

Banerjee Imon; Chen, Matthew C.; Lungren, Matthew P.; Rubin, Daniel L. (2018), *Journal of Biomedical Informatics.*

[4] *Wikipedia. Word2vec* [online] Available at:

https://en.wikipedia.org/wiki/Word2vec

[5] *Towards data science. POS Tagging using CRFs* [online] Available at:

https://towardsdatascience.com/pos-tagging-using-crfs-ea430c5fb78b

[6] *Great Learning. 2020, What is Named Entity Recognition (NER) Applications and Uses?* [online] Available at:

https://www.mygreatlearning.com/blog/named-entity-recognition/

[7] *Analytics Vidhya. 2017. The Essential NLP Guide for data scientists (with codes for top 10 common NLP tasks).*

[online] Available at:

https://www.analyticsvidhya.com/blog/2017/10/essential-nlp-guide-data-scientists-top-10-nlp-tasks/

[8] "*A Survey of Text Similarity Approache*s" Wael H. Gomaa, Aly A. Fahmy. International Journal of Computer

Applications , April 2013 .

[9] "*Evaluation of language identification methods using 285 languages*". Tommi Jauhiainen, Krister Linden, Heidi

Jauhiainen. Proceedings of the 21st Nordic Conference of Computational Linguistics, Gothenburg, Sweden, 23-24 May 2017

[10] *Analytics Insight. 2020 TYPES OF SENTIMENT ANALYSIS AND HOW BRANDS PERFORM THEM*

[online] Available at:

https://www.analyticsinsight.net/types-of-sentiment-analysis-and-how-brands-perform-them/

[11] Monkey Learn. Intent Classification: How to Identify What Customers Want. [online] Available at:

https://monkeylearn.com/blog/intent-classification/

[12] *Medium. Text Classification Algorithms: A Survey.* 2019. [online] Available at:

https://medium.com/text-classification-algorithms/text-classification-algorithms-a-survey-a215b7ab7e2d

[13] Github. Facebook Research FastText. 2020. [online] Available at:

https://github.com/facebookresearch/fastText

[14] *Towards Data Science. Multi-Class Text Classification with Doc2Vec & Logistic Regression.* 2018.

[online] Available at:

https://towardsdatascience.com/multi-class-text-classification-with-doc2vec-logistic-regression-9da9947b43f4#:~:text=Doc2vec%20is%20an%20NLP%20tool,generalizing%20of%20the%20word2vec%20method.&text=Distributed%20Representations%20of%20Sentences%20and,classification%20with%20word%20embeddings%20tutorial

[15] *Orange3 Text Mining. Concordance.* [online] Available at:

https://orange3-text.readthedocs.io/en/latest/widgets/concordance.html

[16] *Monkey Learn. Text Processing: tools, methods, and applications.* [online] Available at:

https://monkeylearn.com/blog/text-processing/

[17] *Monkey Learn. Monkey Learn.* [online] Available at:

https://monkeylearn.com/blog/what-is-tf-idf/#:~:text=TF%2DIDF%20is%20a%20statistical,across%20a%20set%20of%20documents.

[18] *Neptune Blog. Data Augmentation in NLP: Best Practices From a Kaggle Master* [online] Available at:

https://neptune.ai/blog/data-augmentation-nlp

[19] *Insights Thematic. 5 Text Analytics Approaches: A Comprehensive Review* [online] Available at:

https://getthematic.com/insights/5-text-analytics-approaches/

[20] *Analytics Vidhya. Top 6 Regression Algorithms Used In Data Mining And Their Applications In Industry* [online]

Available at:

https://analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/

[21] *Analytics Vidhya. A Comprehensive Guide to Understand and Implement Text Classification in Python* 2018

[online] Available at:

https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/

[22] "*A Survey Paper on Text Mining - Techniques, Applications And Issues*", Mrs.B.Meena Preethi , Dr.P.Radha,

Government Arts College,Coimbatore, India. IOSR Journal of Computer Engineering.

[23] "*A Survey on Techniques in NLP*" Nihar Ranjan, Kaushal Mundada, Kunal Phaltane, Saim Ahmad. International

Journal of Computer Applications, 2016.

[24] *Medium. Dealing with data preprocessing problems.* 2018. [online] Available at:

https://medium.com/@limavallantin/dealing-with-data-preprocessing-problems-b9c971b6fb40

[25] *Monkey Learn. Major Challenges of Natural Language Processing (NLP) for AI.* [online] Available at:

https://monkeylearn.com/blog/natural-language-processing-challenges/