Seema Pal, 2024, 12:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

A Survey on Digital Text Content Classification Features and Techniques

Seema Pal, Professor Sumit Sharma
Department of Computer Science and Engineering,
Vaishnavi Institute of Technology and science Bhopal, MP, India

Abstract- There has been a dramatic increase in the volume of documents and texts in recent years. This needs a more refined machine learning techniques for their proper classification in a wide variety of contexts. In the field of natural language processing, many machine learning techniques have performed better than human experts. To be effective, these learning algorithms need to be able to comprehend not only simple models, but also non-linear connections hidden within the data. This study provides an in-depth analysis of the various document categorization methods currently available, allowing for the precise identification of document classes. The importance of preprocessing in text mining for feature creation was also highlighted. The paper contains extensive background research. Concerns unique to the field were also addressed.

Keywords- Data Mining, Text Mining, Document Analysis, Text Preprocessing, and Related Terms

I. INTRODUCTION

Human civilization can only progress as long as people are able to communicate with one another. When this paper can communicate with one another, this paper can work together to complete challenging jobs. Over the years, humans have developed and refined the alphabet, writing, spelling standards, and other aspects of the written language. That's why this paper live in an era where books, articles, libraries, data corpora, and the entire World Wide Web put an infinite amount of knowledge within our reach. The sheer volume of information that can be found on the Internet, company intranets, and newswires can be intimidating. While there is an ever-increasing trove of knowledge at our fingertips, the human mind can only take in so much before it becomes overwhelmed [1].

The current proliferation of digital document archives calls for innovative approaches to automatically cataloguing, indexing, searching, and navigating huge datasets. New methods for discovering word patterns in document collections utilising hierarchical probabilistic models have been created based on current research in machine learning and statistics. Topic models are the term for this type of model. Hierarchical probabilistic models, for example, are easily generalizable to other types of data; topic models have been used to analyse things other than words, including images, biological data, and survey information and data [2]. Discovering word-use patterns and learning how to link publications with comparable patterns is the primary value of topic modelling. Since texts are mixes of topics—where a topic is a probability distribution over words—topic models are a useful concept for working with such data. Topic model, to put it another way, is a document-based generative model. A straightforward probabilistic method for document generation is outlined.

By analysing the textual data of various types of documents like books, websites, emails, reviews, reports, and product descriptions, text mining is a

cutting-edge method for extracting knowledge from document collections. Both the information providers and the file types used might vary [3, 4]. This is how databases classify their textual resources:

In contrast to unstructured data, structured data is neatly ordered in predetermined rows and columns. Information such as street addresses, biological statistics, and phone numbers are included.

The term "unstructured data" refers to information that lacks a standard format. There might be comments or ratings culled from many online sources. It frequently incorporates acoustic and visual components.

Data that combines elements of both unstructured and structured forms is called semi-structured data, and its name gives away its definition. Popular semi-structured file formats include JSON and XML, which, although being clearly defined, lack the requisite structure for use in a relational database.
Text mining "turns text into numbers" so that the technique may be used regardless of the original data's format. Knowledge of how to utilise and combine approaches for managing text, from single words to documents to complete document databases, is required for converting text into a structured, numerical format and using analytical algorithms [5].

II. TEXT PRE-PROCESSING TECHNIQUES

Preprocessing strategy assumes a significant job in content mining methods and applications. It is the initial phase in the content mining process. This work examine three key strides of preprocessing in particular, tokenization, stop words expulsion, stemming.

1. Tokenization
This procedure split the grouping of strings into words. It expels all the accentuations from the content information and gives expressions of content which is called tokens [5].
Fast Miner gives you three different ways to split the data. The most common way is to use the default method, which is called "non-letter character." This method splits the data based on non-letter characters like spaces, commas, full stops, and so on. The next mode is "determine character," in which you can specify characters based on how the sentence is broken up into tokens. The third mode is "standard articulation," in which a regular phrase is given to break up the sentence into tokens.

2. Stop Words Elimination
Stop words are a division of natural language. The thought process that stop-words ought to be expelled from a book is that they make the content look heavier and less significant for experts. Evacuating stop words decreases the dimensionality of term space. The most widely recognized words in content archives are articles, relational words, and professional things, and so forth that doesn't give the significance of the records. These words are treated as stop words. Model for stop words: the, in, an, a, with, and so on. Stop words are removed from documents because those words are not measured as keywords in text mining applications [6].

3. Stemming
This method is used to identify the root/stem of a word. For example, the words connect, connected, connecting, connections all can be stemmed to the word “connect” [7]. The purpose of this strategy is to expel different postfixes, to diminish the quantity of words, to have precisely coordinating stems, to spare time and memory space.

4. POS Tagging
Tagging in characteristic language preparing (NLP) alludes to any procedure that appoints certain marks to phonetic units. It signifies the task of grammatical feature labels to writings. A PC program for this purpose is known as a tagger. Grammatical form tagging incorporates the way toward allotting one of the grammatical features to the given word. For example, the english word rust for instance is either a verb or a noun. Some of dictionary are available such as Maxent Tagger from Stanford CoreNL [8].

Features of Text Mining

TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

$$TF(t) = \frac{Number\ of\ Times\ term\ t\ Present\ in\ Document}{Total\ Number\ of Terms\ in\ \ Document}$$

IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus this paper need to weigh down the frequent terms while scale up the rare ones, by computing the following:

$$IDF(t) = log\left(\frac{Total\ Number\ of Document}{Number\ of\ \ Document 10aving\ term\ t}\right)$$

TF-IDF (Term Frequency Inverse Document Frequency) This term composed by two terms: the first computes the normalized Term Frequency (TF). The number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

TF-IDF = TfxIDF

III. RELATED WORK

Lawrence Reeve in [6] has Semantic annotation deals with enriching texts with pointers to knowledge bases and ontologism. Past work for the most part centered around connecting notices of ideas and cases to either semantic vocabularies like Word Net, or Wikipedia-based information. DBpedia was for instance used to consequently remove theme marks by connecting the inalienable points of a content to ideas found in DBpedia and mining the subsequent semantic subject charts. They found this is a superior approach than utilizing content based strategies. Feeling examination, then again, manages discovering suppositions in content. Most research has been performed on obviously obstinate messages, for example, item or motion picture re-sees [5], rather than daily paper writings which are accepted to be less stubborn.

Elad Segev in [7] which visualized distinguishing predispositions when investigating Israel, found that news tweet / comments are to a great extent basic and negative towards Israel. More subjective examinations were performed, for example, the talk which uncovered complexity designs that give proof to ideological contrasts amongst neighborhood and universal press scope. These examinations either center around a specific occasion or subject or utilize content characterization keeping in mind the end goal to characterize themes, and regularly require a forthright meaning of points as well as physically clarified preparing information. In this work, rather, this work utilize semantic web innovations to semantically comment on newswire message, and build up a completely programmed pipeline to discover questioned themes by utilizing notion examination strategies.

Document Classification

Ying, Y., et al. [8] (2017) used abstractive text summarization to extract essential terms. During text extraction, they used a graph-based approach to prioritise key terms. They chose an abstracted text extraction strategy to connect the keywords while ignoring the sentencing implications.

They recommended specific metadata links between standard terms and their associated sentences. They started by grouping several publications into clusters, which then displayed the key issues of the articles. They devised three distinct criteria and demonstrated that their work was superior to simply extracting important phrases.

To address the earlier limitations of manual text processing, Gupta, A. et al. [9] (2018) examined radiological reports using abstractive

summarization and Clustering. They enhanced keyword extraction and text grouping in dictionary- and rule-based techniques. The cluster link analysis was missed by the named entity recognition procedure in previous processing methodologies. They used an unsupervised approach to extract named entity relationships with no prior knowledge. They used parse trees to cover text processing requirements and then used distributed semantics to connect them.

Moradi, M. [10] (2018) employed extractive text summarization to condense the text by removing unnecessary information. Clustering & Itemset-mining Biomedical Summarizer was the term given to this technology, which used text summarization to summarise biomedical data (CIBS). This method extracted biomedical concepts from the text input. By applying the itemset mining algorithm to reduced text, the ideas represented the major subjects, and CIBS placed sentences into the clusters' relevant set.

KUSH is a text summary approach introduced by Uçkan et al. [11] (2020). This method determines the largest number of non-overlapping abstractive summarization sets possible. To determine the context of various paragraphs in unstructured text texts, they labelled these sets as nodes. They concentrated their efforts on creating a logical text visualisation. Abstractive summarization was combined with set theory and graph visualisation tools in their proposed KUSH technology.

In [12], the authors suggest, analyse, and compare a multi-modal binary classication network that uses text and pictures on digital document pages to state-of-the-art baseline methods. From the image of the input document, both image and textual features are extracted at the same time by sending them through the Visual Geometry Group 16 - Convolutional Neural Network (VGG16-CNN) and the pre-trained Bidirectional Encoder Representations from Transformers (Legal-BERT base) model through transfer learning. Both features are finally combined and sent through a fully linked layer of Multi Layered Perceptron (MLP) to get the binary classification of the pages as the First Page (FP) and the Other Page (OP).

In [13] authors proposed a novel framework that can evaluate the quality of documents in terms of consistency. This study uses a binary classification task to describe how to find low-quality documents. This task can measure how well the documents have consistent content. In particular, this study makes the problem less difficult by thinking of each line or paragraph as a node. Then, a certain text is looked at as a network of points. This study shows how the super node in a network is defined and how it can be used to tell if a document is consistent or not.

IV. CHALLENGES IN TEXT SENTIMENT ANALYSIS

1. Detection of Spam and Fake Reviews

The Internet has both real and fake content. For strong Sentiment grouping, this junk content should be removed before being handled. This should be possible by telling copies apart, finding oddities, and taking the commentator's reputation into account [1].

2. Limitation of Classification Filtering

There is a limit to sorting things in order and choosing what is most popular at the same time. For better labelling of ideas, this restriction should be loosened. The danger of the channel bubble [11] is that it gives too many input sets, which leads to a fake list of opinions.

Domain-Independence

The space-independent quality of emotion words is the most important test that opinion mining looks for. One set of features may work well in one place but very poorly in another.

Natural Language Processing Overheads

The general language baggage, like ambiguity, co-reference, implicitness, derivation, and so on, made it harder to investigate how people felt.

The sentiment given on twitter is hard to grasp as it comprises of poor words, absence of capital letters,

spelling botches, no appropriate accentuations, and syntactic blunders, etc.

After some time, the analyst's feelings change. People's mental states are studied to see how they change over time in [9]. The search looks for places where the researcher's personality is clearly shown, either by looking through a list of personality traits or by writing it as a free-form sentence.

V. CONCLUSION

Multi-label classification of text (TC) is an important part of TC because text classification is one of the most important tasks in natural language processing (NLP). As the field moved away from standard machine learning and towards deep learning, MLTC models that work well have grown over time. But at the moment, numerous linked studies have different foci. In this study, this paper have compiled the findings of a number of researchers that have studied the problem of text document categorization using machine learning techniques. It was found that use of machine leaning for classification has improve the work efficiency. Apart from this paper has find that term and pattern feature based learning is mostly adapt by the scholars. In future scholars can develop model that provide multiclass prediction of text file without prior information.

REFERENCES

1. Feldman, R. & Sanger, J. (2007). The text mining handbook. Advanced Approaches in Analyzing Unstructured Data.Cambridge: Cambridge University Press.
2. K.;Meimandi, K. J.;Heidarysafa, M.;Mendu, S.;Barnes, L. & Brown, D. (2019). Text Classification Algorithms: A Survey. Information.
3. Miner, G.;Delen, D.;Elder, J.;Fast, A.;Hill, T.&Nisbet, R. & Balakrishnan, K. (2012). Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.Waltham: Academic Press is an imprint of Elsevier.*** (2020).
4. Natural Language Processing (NLP). IBM Cloud Education:https://www.ibm.com/cloud/learn/natural-language-processing.
5. Singh, J. & Gupta, V. (2016). Text Stemming: Approaches, Applications, and Challenges. ACM Computing Surveys, pp. 1-46.
6. Lawrence Reeve and HyoilHan.”Survey of semantic annotation platforms ACM symposium on Applied computing”,. ACM, 2005, pages 1634-1638.
7. EladSegev and Regula Miesch. ”A systematic procedure for detecting news biases: The case of israel in european news sites. International Journal of Communication”, 2011.
8. C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, ``The Stanford CoreNLP natural language processing toolkit,'' in Proc. 52nd Annu. Meeting Assoc. Comput. Lin-guistics, Syst. Demonstrations, 2014, pp. 55_60.
9. Lawrence Reeve and Hyoil Han.”Survey of semantic annotation platforms ACM symposium on Applied computing”,. ACM, 2005, pages 1634-1638.
10. Elad Segev and Regula Miesch. ”A systematic procedure for detecting news biases: The case of israel in european news sites. International Journal of Communication”, 2011.
11. Y. Ying, T. Qingping, X. Qinzheng, Z. Ping, and L. Panpan, ''A graphbased approach of automatic keyphrase extraction,'' Procedia Comput. Sci., vol. 107, pp. 248–255, 2017.
12. A Gupta, I. Banerjee, and D. L. Rubin, ''Automatic information extraction from unstructured mammography reports using distributed semantics,'' J. Biomed. Informat., vol. 78, pp. 78–86, Feb. 2018.
13. M. Moradi, ''CIBS: A biomedical text summarizer using topic-based sentence clustering,'' J. Biomed. Informat., vol. 88, pp. 53–61, Dec. 2018.
14. T. Uçkan and A. Karci, ''Extractive multi-document text summarization based on graph independent sets,'' Egyptian Informat. J., vol. 21, no. 3, pp. 145–157, Sep. 2020.
15. A. Guha, A. Alahmadi, D. Samanta, M. Z. Khan and A. H. Alahmadi, "A Multi-Modal Approach to Digital Document Stream Segmentation for

Title Insurance Domain," in IEEE Access, vol. 10, pp. 11341-11353, 2022.

16. D. Jung, M. Kim and Y. -S. Cho, "Detecting Documents With Inconsistent Context," in IEEE Access, vol. 10, pp. 98970-98980, 2022.