INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

OPEN ACCESS

Received: 30-07-2022

Accepted: 28-09-2022

Published: 10-11-2022

Citation: Shilpa BL, Shambhavi BR (2022) Structuring of Unstructured Data from Heterogeneous Sources. Indian Journal of Science and Technology 15(41): 2188-2193. https://doi.org/10.17485/IJST/v15i41.1566

* **Corresponding author**.

shilpabl@gsss.edu.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Structuring of Unstructured Data from Heterogeneous Sources

B L Shilpa[1]*, B R Shambhavi[2]

1 Assistant Professor, Department of ISE, GSSSIETW, Mysuru, Karnataka, India
2 Associate Professor, Department of ISE, BMSCE, Bengaluru, Karnataka, India

Abstract

Objectives: To develop a new data gathering processing under Big Data Perspectives. To convert unstructured text data into structured format by not missing out any text data available. **Methods:** The unstructured data is pre-processed using modified stemming and tokenization. From the stemming output, the proposed Term Frequency-Inverse Document Frequency (TF-IDF) and N-gram features are derived. Unstructured data is considered from multiple sources like twitter, consumer complaints and news blog. **Findings:** The proposed model with extant TF-IDF features has exposed relatively high Mean Average Error (MAE) value which is 1.4325 when compared to the proposed model without optimization to be 0.5197. **Novelty:** The novelty of the research work is of the stemming process where dictionary checking process is added and the improved feature extraction, interclass dispersion coefficient is computed in TF-IDF features.

Keywords: Natural language processing; Structured data; Unstructured data; Big data; Feature extraction

1 Introduction

Unstructured data are those types of data that do not have any predefined format or structure. There are no rules to create Unstructured Data. Normally, an Unstructured Data is in text format, such as open-ended surveys or answers to social media conversations, but can also be non-textual in nature, such as an Image, Video or an Audio file.

Since there is a rapid increase in the usage of internet, there is a rise in the production of unstructured data at an alarming rate. This unstructured data that is being produced is estimated to be around 90% which is giving an indication to the organizations to do something quickly about these data that is being produced. Though structured data is important for organizations, unstructured data is like hidden gem when analyzed properly it produces drastic results to the organizations[(1)]. They provide a wealth of insights that cannot be explained by statistics or numbers alone.

This structured, semi-structured and unstructured data are so huge in quantity that they all fall under the umbrella of 'Big Data'[(2)].

All these three types of data can provide substantial insights, but it's important for one to first know the 'type' of data, in order to get the insights needed.

Although it contains numbers, statistics and facts, an Unstructured Data is usually composed of methods which are text-rich or difficult to analyse.

For example, a social media post can include opinions, discussed topics and feature recommendations. However, it is difficult to process this information in large quantities. Therefore, in order to reveal practical insights, certain information first needs to be extracted, categorized, and analysed.

Figure 1 illustrates how unstructured data can be textual or non-textual and take the shape of text, audio, and images. This information is an important component of an organization's knowledge base and must be appropriately handled for long-term use. It offers a significant window of opportunity for profitable economic outcomes. Effective unstructured data management can increase revenue, profitability, and potential while decreasing risks and expenses. Unstructured data offers insights into consumers' motivations, future goals, and potential issues in contrast to structured transaction data, which reveals what customers did.

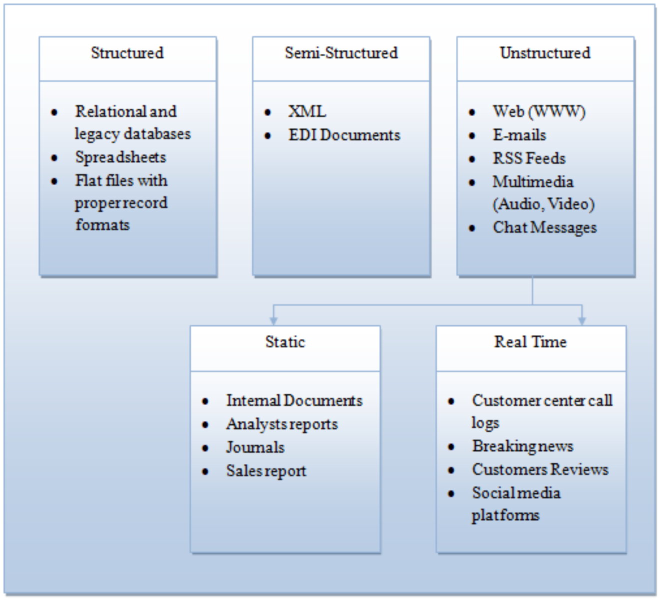

Fig 1. Types of Data

The proposed method creates a system that can quickly extract the necessary information from multiple web pages or URL and organize it into a structured format. From this structured data, any user requested data can be retrieved and used for further processing. We use Python's NLTK and few other packages to accomplish this. Structuring of unstructured data helps in analyzing that 90% of data which is going waste for organizations. In this work the sole concentration is on structuring unstructured data related to stock market only. This was chosen because lots of unstructured data related to stock market go unutilized.

Approaches based on pattern discovery can be used to extract structured data from semi-structured Web content. Recent iterations of these methods focus on extracting patterns from Web pages devoid of user-labeled samples utilizing a variety of pattern discovery techniques, such as radix trees, multiple string alignments, and pattern matching algorithms[3]. These data extractors can be applied to pages from the same Web data source that have not yet been seen but will not be able crawl from other web page, whereas the web crawler designed here crawls unstructured data from multiple sources at a single time.

In paper[4] the authors are trying to predict the sentiment from the whole lot of unstructured text available. As most of the existing system work on supervised learning technique, which requires a labelled dataset and building a labelled dataset for every domain is an impossible task. In this paper a cross domain label set is used. Here training is done using one domain data set and performance evaluation is done on another domain. Cross-domain sentiment analysis has shown limited performance improvement as it suffers with the problem of large number of unseen words (out-of-the-vocabulary). This drawback has been overcome in our proposed technique as we have added a dictionary during feature extraction which has resulted in better performance.

The paper[5] presents a classification model which supports both the generality and the efficiency. The generality is supported by following a logical sequence process and classifying the unstructured text documents step by step. This paper shows the importance of unstructured text documents classification. Here based on the documents contents they are classified into predefined categories by a set of phases and each phase is carried out using different techniques. In the Proposed technique instead of crawling all the data and the classify it, we have designed the crawler in such a way that based on certain key words given related to the domain the unstructured data that is only relevant are crawled. Hence there is no need for any classification model later during preprocessing step.

Here in paper[6] various unstructured data related to stock market is considered to do the prediction of stock market price movement. The unstructured data considered are from various sources like twitter, Facebook, online news, google trend and forum discussion. This paper introduces a Spark[7] NLP- based text preprocessing pipeline for removal of the noisy data and features are extracted using TFIDF[8] method. Whereas in our proposed technique instead of using the existing TFIDF we add dictionary to the same and use which gave a better result. Here two library Textblob and Vader are used for performing the sentiment analysis on the unstructured data. Whereas the proposed method uses the Natural Language Toolkit (NLTK) conglomerate of Python and its libraries. It's also making use of Python standard library urllib, which contains tools for working with URLs is used to perform web scraping or crawling of unstructured data from multiple sources. The proposed method creates a system that can quickly extract the necessary information from multiple web pages or URL and organize it into a structured format.

2 Methodology

The popularity of internet has resulted in an explosion of information being produced online and it has become extremely difficult for organizations to find which information is useful and how to extract value out of this information or data. Also, this information is not present in any central repository that we apply some queries and process it. It needs to be extracted from multiple sources like twitter, Facebook, online news, Forum discussion etc. Because of this bombarding of information on the web it is very difficult to understand which information should be extracted and which contains potentially useful knowledge. Extraction of relevant information from this unstructured data is not only challenging it is also time consuming because of the huge volume. Hence, conversion of this unstructured data to structured data has become very much necessary for organizations especially in the field of Finance and Banking.

The developed model for predicting stock price includes the following phases.

- Initially, the web crawler crawls the unstructured data from multiple sources and stores it for further processing
- Next, pre-processing is performed, during which the sentiment data (text) is subjected to modified stemming (text) and tokenization (integer).
- Further, the features like BOW, proposed TF-IDF and N-gram features are derived from the modified stemming output.

Figure 2 shows the overall depiction of the suggested system to convert unstructured data into structured format.

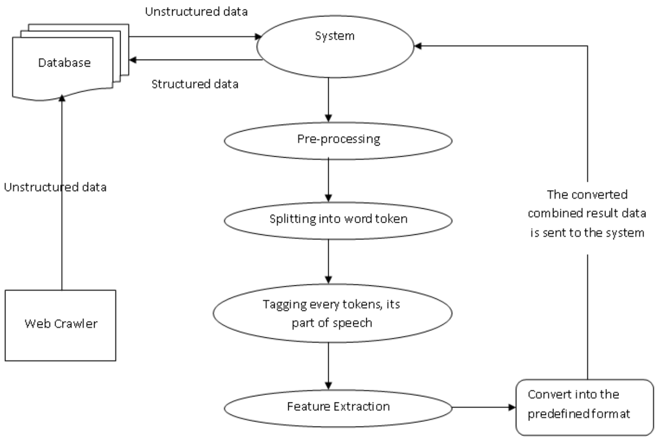
Unstructured data
System
Database
Structured data
Pre-processing
Unstructured data
Splitting into word token
The converted combined result data is sent to the system
Web Crawler
Tagging every tokens, its part of speech
Feature Extraction
Convert into the predefined format

Fig 2. Pictorial representation of Proposed System

Improved Stemming[9]: It is a technique to eliminate suffix of word and it aids in minimizing the needed number of computations. Stemming is deployed in diverse appliances like information retrieving systems. Moreover, it is exploited in domain analysis for determining the domain vocabulary. In this work, improved stemming process is performed. However, there is no checking of dictionaries. This has an impact on all words that goes to stemming process. Here, we have added dictionary checking process in first step. After that based upon concurrence analysis, the stemming process is done.

For two terms, *a* and *b*, the expected mutual information *EM* is modelled as in Eq. (1), wherein, n_{ab} refers to count of times *a* and *b* co-occur in a text window of fixed size, n_a and n_b refers to count of occurrences of *a* and *b* in corpus, *EM(a,b)* refers to expected count of occurrences of *a* and *b*, *k* refers to constant based upon corpus and window size. Here, *k* is computed as shown in Eq. (2).

$$EM(a,b) = \max\left(\frac{n_{ab} - EM(a,b)}{K * n_a + n_b}, 0\right) \tag{1}$$

$$K = \frac{\sum n_{ab}}{\sum n_a n_b} \tag{2}$$

Tokenization: This process converts the text into tokens prior to transferring to vectors. It is simple to filter the unnecessary tokens. In addition, tokenization[10] is the procedure of splitting the unprocessed text into small chunks. The raw texts are broken down into sentences, words termed as tokens during tokenization. Subsequently, the tokens help to know the context for NLP and aids in understanding the meaning of text via examining the sequence of words.

2.1 Feature Extraction

2.1.1 BOW Features

The text is converted into a bag of words in BOW[11]. With size m x n, the feature matrix is created. Here, m exposes the sentence count in corpus and n exposes count of unique words. The derived BOW oriented features is signified as FT^{BOW}

2.1.2 TF-IDF Features

TF–IDF[8] is a significant format of text demonstration and includes longer history amongst 3 well known depiction techniques. It depends upon the BOW method, where a text is characterized by a compilation of words deployed in the document. The constraint TF_{ij} is describes as the count of times word i appear in document j; the better the value, the more noteworthy the word will be. The constraint DF_i signifies document count, where word *i* appears once; the better the value, the more frequent the word is. If word i is significant for document j, it must comprise a superior TFij and lesser DFi. Conventionally, TF-IDF features are expressed as in Eq. (3), where, TFij signifies the count of times word i appear in document j; m signifies document count in collection, ni signifies entire count of documents, where features appear. As per improved concept, TF-IDF features are expressed as in Eq. (4), in which D_i refers to inter class interclass dispersion coefficient and is computed as shown in Eq. (5). In Eq. (6) n refers to the count of classes and F (t, i) refers to the count of document with t and it belongs to same class where term belongs to.

$$f^{TF-IDF} = TF \times IDF_{ij} = TF_{ij} \times \log\left(\frac{M}{n_i} + 0.01\right) \tag{3}$$

$$f^{TF-IDF} = TF \times IDF_{ij} \times D_i \tag{4}$$

$$D_i = \left(\prod_{i=1}^{n}(F(t,i) - \operatorname{avg}(F(t,i)))^2\right)^{\frac{1}{2}} \tag{5}$$

$$\operatorname{avg}(F(t,i)) = \frac{1}{n}\sum_{i=0}^{n} F(t,i) \tag{6}$$

The extracted improved TF-IDF based features are denoted as $FT^{ITF\text{-}IDF}$

2.2 n-gram Features

An n-gram model[12] is defined as “a method of including sequences of words or characters that permits us to maintain richer pattern discovery in text, i.e. it attempts to captivate patterns of sequences (words or characters subsequent to one another) while being responsive to appropriate relations (words or characters subsequent to one another)”. The extracted n-gram based features are denoted as $FT^{n\text{-}gram}$

The features of sentiment data are indicated as FT^{SD}, and it is shown in Eq. (6).

$$FT^{SD} = FT^{n-gram} + FT^{ITF-IDF} + FT^{BO\,W} \tag{7}$$

3 Results and Discussion

The deployed approach for converting unstructured data into structured format was done in "PYTHON". The data set was collected from multiple sources[13] like news data, twitter data, complaints forum and various related blogs. All the unstructured data collected was related to stock of two companies mainly, Reliance Communications and Relaxo Footwear. Since the processed data was further used for prediction of stock prices[14].

The proposed optimized TF-IDF technique is applied to improve the quality of text data conversion. The analysis is done on the existing System without optimization by tuning the weights to provide better results. The analysis is done by varying the learning percentage that ranges from 60 to 90 with respect to metrics like Mean Average Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

The Mean Average Error (MAE) of the proposed optimized TF-IDF technique is obtained to be 1.43 when compared to the existing TF-IDF with any optimization which has an MAE of 0.52. We observe that in MAE the proposed method gives 89.6% of efficiency when compared to the existing system without optimization. This is because of the interclass dispersion coefficient computation that is done in the TE-IDF method. With respect to the Mean Squared Error (MSE) value it gives an effienecy of 51.66 % and lastly it was compared with the Root Mean Squared Error (RMSE) value where we obtained an effiency of 70% when compared to the exisiting one.

Table 1 illustrates the study of, proposed model with extant TF-IDF features, consequently, study is made on diverse metrics like MAE, RMSE and MSE. On noticing the results, the proposed TF-IDF features have attained finest values than the proposed model without optimization. Moreover, the developed model has exposed relatively high MAE values than the proposed model without optimization. This demonstrates that none of the unstructured text data has been left without pre-processing. Complete data that is crawled is converted into structured format.

Table 1. Analysis on existing methods as well as proposed optimization theory

Metrics	MAE	MSE	RMSE
Proposed without optimization	0.51979	0.607331	0.470995
Proposed with extant TF-IDF	1.432516	1.283011	1.132581
Optimized TF-IDF Method	1.896817	1.511884	1.91622

4 Conclusion

This work developed a new unstructured to structured text conversion model, where unstructured text data related to stock market was considered. The data gathering was done using a web crawler to crawl the data from multiple sources. This crawler was able to crawl the data from multiple sources at a time where as the existing one was not able to do so. The pre-processing was done using modified stemming and tokenization. From which the required features are derived. The proposed method gives 89.6% of efficiency when compared to the existing system without optimization. This converted structured data is further used to predict the stock price conserving the historical data. This system is not just limited to the conversion of unstructured data to structured data of stock market, it can be expanded to many other domains like banking, Finance, education etc.

References

1) Kumar A, Dabas V, Hooda P. Text classification algorithms for mining unstructured data: a SWOT analysis. *International Journal of Information Technology*. 2020;12(4):1159–1169. Available from: https://doi.org/10.1007/s41870-017-0072-1.
2) Giudice PL, Musarella L, Sofo G, Ursino D. An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake. *Information Sciences*. 2019;478:606–626. Available from: https://doi.org/10.1016/j.ins.2018.11.052.
3) Zaman G. Information extraction from semi and unstructured data sources: A systematic literature review. *ICIC Exp Lett*. 2020;14:593–603. Available from: http://www.icicel.org/ell/contents/2020/6/el-14-06-09.pdf.
4) Kumari S, Agarwal B, Mittal M. A Deep Neural Network Model for Cross-Domain Sentiment Analysis. *International Journal of Information System Modeling and Design*. 2021;12(2):1–16. Available from: https://doi.org/10.4018/IJISMD.2021040101.
5) Mowafy M, Rezk A, El-Bakry H. An efficient classification model for unstructured text document. *American Journal of Computer Science and Information Technology*. 2018;6(1):16. Available from: https://doi.org/10.21767/2349-3917.100016.
6) Chen L, Kong Y, Lin J. Trend Prediction Of Stock Industry Index Based On Financial Text. *Procedia Computer Science*. 2022;202:105–110. Available from: https://doi.org/10.1016/j.procs.2022.04.014.
7) Kocaman V, Talby D. Spark NLP: Natural Language Understanding at Scale. *Software Impacts*. 2021;8:100058. Available from: https://doi.org/10.1016/j.simpa.2021.100058.
8) Kim D, Seo D, Cho S, Kang P. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. *Information Sciences*. 2019;477:15–29. Available from: https://doi.org/10.1016/j.ins.2018.10.006.

9) Khyani D, Siddhartha BS, Niveditha NM, Divya BM. An Interpretation of Lemmatization and Stemming in Natural Language Processing. *Journal of University of Shanghai for Science and Technology*. 2021. Available from: https://jusst.org/an-interpretation-of-lemmatization-and-stemming-in-natural-language-processing/.

10) Solangi YA, Solangi ZA, Aarain S, Abro A, Mallah GA, Shah A. Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis. *2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*. 2018;p. 1–4. Available from: https://doi.org/10.1109/ICETAS.2018.8629198.

11) Pimpalkar AP, Raj RJR. Influence of Pre-Processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*. 2020;9(2):49–68. Available from: http://digital.casalini.it/5010980.

12) Awwalu J, Bakar AA, Yaakub MR. Hybrid N-gram model using Naïve Bayes for classification of political sentiments on Twitter. *Neural Computing and Applications*. 2019;31(12):9207–9220. Available from: https://doi.org/10.1007/s00521-019-04248-z.

13) Atay M, Kalayci M, Apik H, Aybar V, Serin F, Akyuz AO. An Approach to Analyzing the Layout of Unstructured Digital Documents. *2022 30th Signal Processing and Communications Applications Conference (SIU)*. 2022;p. 1–4. Available from: https://doi.org/10.1109/SIU55565.2022.9864787.

14) Lang HX, Li YY, Wang Y, Wang H, Dong J. An Automatic Topic-oriented Structured Text Extraction Method based on CRF and Deep Learning. *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. 2022;p. 1408–1413. Available from: https://doi.org/10.1109/CSCWD54268.2022.9776155.