A Survey on Text Mining and Sentiment Analysis for Unstructured Web Data

[1] Nikhil R, [2]Nikhil Tikoo, [3]Sukrit Kurle, [4]Hari Sravan Pisupati, [5]Dr. Prasad G R
[1-4] U.G. Scholar, Department of Computer Science and Engineering, BMS College of Engineering, Bangalore, India.
[5] Associate Professor, Department of Computer Science and Engineering, BMS College of Engineering, Bangalore, India.

Abstract: Unstructured data refers to information that doesn't have a pre-defined data archetype. Unstructured information is typically textual data, but may also contain numerical data, and factual details. This results in data that is obscure, irregular and ambiguous, thus making it difficult to analyse using conventional computing means. Much of the data in the web, in the form of blogs, news, social media platforms is unstructured. But they serve as a potential vast source of information, if processed efficiently. In this paper, the basics of harnessing unstructured data from the web and the techniques to process it are discussed. The concepts of web crawling, text mining and natural language processing are discussed in brief, to give an outline of how web data is processed and analysed. Sentiment Analysis, which is a major aspect of present day NLP, is also described, along with issue of mining from Twitter, which has emerged as the most important data source for NLP in the recent past. The paper concludes with a brief outline of the use of web data mining and analysis, and the potential for future growth in the field.

Keywords **– Data Mining, Natural Language Processing (NLP), Sentiment Analysis, Text Mining, Web Crawling.**

1. INTRODUCTION

Data mining refers to the process of extraction of knowledge from large sets of data. It is also aptly termed as 'Knowledge Discovery in Databases (KDD)'. This trend of analysis and extraction of information from data sets has spawned a whole new field of study and an array of techniques to do so.

Data collection and storage technology has made it possible for organizations to gather and maintain huge amounts of data at low costs. The generic activity and the overall objective of data mining is to exploit this stored data in order to extract useful information. Many organizations maintain 'data warehouses', which contain historical data maintained for the sole purpose of analysis and decision-making.

Occasionally, information is extracted from resources distributed over the web. This can be achieved with the help of a Web Crawler to parse and retrieve web pages. Often collected in an unstructured form, this data must be converted into a structured format that is suitable for processing. Techniques such as Natural Language Processing (NLP), text-analytics and Latent Semantic Analysis (LSA) provide different methods to interpret this information.

Once analysed, user-designed GUI's (Graphical User Interfaces) display the results of the analysis in an interactive, user-friendly manner which allows for easy, quick understanding. Organizations also build and maintain GUI's to display the analysed data from their data warehouses.

2. SURVEY

2.1. *WEB CRAWLING*

The World Wide Web and its contents serve as the most important source of information nowadays. Because much of its contents is *unstructured*, it has to be efficiently mined and analysed to extract useful information from it. Web mining has today emerged as one of the most important research areas in computer science.

According to Subashini S and Mahesh T.R [1], *Web Mining* is the application of data mining processes to extract definite knowledge from web data, and can be categorized as – *content mining*, *structure mining* or *usage mining.* Some techniques used for this purpose include pre-processing, detection and filtering – using classification models, understanding user behaviour through user profiles etc. Some of the most important applications in this field are displayed by the work of Amazon.com, Google, AOL and Yahoo. For example, Amazon.com uses web mining to perform cross-referencing of items and provide recommendations. In the present scenario, much emphasis is being placed on optimizing web services, fraud and online crime analysis. Thus, web content security and access protection are major issues of importance in this field.

Web Crawling is the first step towards web mining. Subhendru Kumar Pani et al. [2] aptly describe a crawler as a program that retrieves web pages, commonly for use by search engine or a web cache, by searching for deviations to web pages, web text and HTML tags, thereby indexing the web. The crawler performs the initial step in all types of web mining- *data, content, structure and usage mining*. According to Gautam Pant et al. [3], crawlers are programs that exploit the graph structure of the Web to move from page to page. The crawler maintains a list of unvisited URLs, called the frontier. Each crawling loop involves picking the next URL to crawl from the frontier, summoning the page which corresponds to the URL through HTTP, parsing/resolving the summoned page to extract the URLs and application-specific information, and finally adding the unvisited URLs to the frontier. The crawler continues this cycle as long as the frontier is not empty.

There are multiple crawling strategies prevalent. Pavalam S M et al. [11] describe five main types of crawling strategies- *Breadth First Search, Depth First Search, Page Rank, Genetic and Naive Bayes Classification.* Though each has its pros and cons, the Genetic Algorithm has the maximum advantages, due to its iterative selection strategy. Crawlers can also be classified based on their behaviour. S.S. Dhenakaran and K. Thirugnana Sambanthan [12] categorize the types of crawlers as- Topic Focused Crawlers(which crawl pages pertaining to a specific topic), Path Ascending Crawlers (which ascend to every path in each URL), Parallel Crawler (which run multiple processes in parallel), and Adaptive Crawlers (which incrementally and continually crawl the entire web).

However, there are certain problems that are encountered while crawling the web, which are well described by Carlos Castillo and Ricardo-Baeza-Yates [4]. Some of them are-

- *Variable Quality of Service*: It refers to the challenge of downloading pages from multiple sources in a stream of data that is as uniform as possible.
- Problems with *real time content update* on web pages, and the difficulty for crawlers to be synchronised with these updates.
- *HTTP Implementations*: Addresses issues such as URL's with no extensions, or with ambiguous extensions, range errors, server responses lacking headers, web servers indicating wrong dates, etc.
- *HTML Coding*: These issues include malformed or syntactically relaxed HTML Markup.

The practical problems of web crawling are mostly related to bad implementations of some web servers and web applications. These issues are not visible until a significant number of pages have been downloaded. Hence, the wrong implementations listed above are constraints that the web crawler designer must consider prior to crawling.

2.2. *TEXT MINING, NATURAL LANGUAGE PROCESSING*

After the web crawling phase, the gathered text has to be mined by *text mining*, which involves mining the crawled textual data using NLP (Natural Language Processing) techniques to obtain meaningful, useful information from it. Mahesh T R et al. [5] have described text mining as a highly important and upcoming technique of the KDD process. It is a multidisciplinary field, involving information retrieval, clustering, machine learning and data mining, among others. It can consist of 2 phases-

1. *Text refining*- In this phase, the free-form text documents are transformed into an intermediate form (IF).
2. *Knowledge distillation*– In this phase, patterns and knowledge are extracted from the Intermediate Form.

There are two basic parameters for assessing the quality of text retrieval process:

1. *Precision*: This is the percentage of retrieved documents that are in fact relevant to the query.
2. *Recall*: This is the percentage of documents that are relevant to the query and were in fact retrieved.

There are multiple techniques employed for efficient text mining. One method is described in lucid detail by Giuseppe Carenini et al [6]. According to them, extracting knowledge from evaluative/free-form text involves two tasks- firstly, extracting relevant information from the text, and secondly, presenting it to the user. For the first task, the most important thing is to determine the important features and terms in the text, so that useful knowledge can then be generated by collecting information by features. This is where machine learning techniques become useful. In the paper's approach, a large, unmanaged list of features obtained by association rule mining is taken as the input. An added component is a *user-defined taxonomy of features*, called 'crude features'. The input can be mapped with this taxonomy to remove redundancy and improve organization. *Similarity matching* is used for this purpose. The similarity matching process generates a set of 'merged features', after which error-checking can be done to eliminate inaccuracies.

Natural Language Processing (NLP) is a field which is closely coupled with text mining. NLP techniques and methods are used widely to process textual data and extract meaningful knowledge and understanding from the text. According to Gauri Rao et al [7], the goal of NLP is to enable communication between people and computers without resorting to memorization of complex commands and procedures, i.e. computers will be ingrained with the ability to understand and generate natural language. A complete realization of NLP capabilities will be far-reaching and have hugely significant applications in the science of data mining.

NLP is linked to the field of Information Retrieval (IR), which is the process of extracting relevant information from textual data. Matthew Lease [20] touches upon this association. According to him, NLP can play an important role not just in mainstream IR but also in Text Retrieval (TR), where a pyramid model of Machine Translation is used to perform information retrieval. An important aspect of Information Retrieval is how to process 'conversational' or 'spontaneous' speech (CS) using NLP, and the problems that arise in this regard- mainly problems with the sentential structure of CS. The lack of a proper structure for CS can cause problems when trying to perform segmentation of such text. Thus, research on how to streamline NLP with IR is still being conducted.

Lots of attempts are being made to build a system which can perform full-fledged NLP. One of them is described by Ronan Collobert et al. [15]. They propose a unified neural network architecture and learning algorithm that can be applied to various NLP tasks. To achieve this versatility, instead of applying specific man-made features, the system learns internal representations on the basis of large sets of mostly un-labelled data used as training sets. It results in a freely available tagging system with good performance.

An emerging scenario in text mining and NLP is *mining Big Data*. According to Albert Bifet [9], streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge, thereby allowing organizations to quickly respond to issues which pop up in real time. This is becoming true especially where very large data sets are involved. The three main dimensions in big data stream mining are *accuracy*, *amount of space* (computer memory) necessary and the *time required* to learn from training examples and to predict. A major application of the above is in social networking profiles like Twitter. Twitter data follows the data stream model. However, as data arrives at high speed, algorithms that process them must do so under very strict constraints of space and time. Another challenging task may be the Structured Pattern Classification problem. Patterns are elements of sets endowed with a partial order relation. Xindong Wu et al. [13] also analyse in detail the issues in big data mining, which are mainly- the volume of big data, data distribution, complex data characteristics, requirement of computational-intensive processors, and the data semantics involved with such immense data volumes. Designing mining algorithms while taking into account all these issues is a very difficult task indeed.

2.3. *SENTIMENT ANALYSIS*

Sentiment analysis, which is the technique of analysing a sentence of text to decipher its sentiments, is one field of NLP which is attracting great attention from researchers. News and blogs are usually good sources of data for sentiment analysis, wherein people can express their thoughts and opinions on such forums. Namrata Godbole et al. [18] have researched on this topic. A text corpus of news entities and blogs is taken as the source data set. The processing system includes a sentiment identification phase in the beginning. It identifies the various entities in the text and associates expressed opinions with each relevant entity. The second phase involves sentiment aggregation and scoring, where each entity is scored relative to others in the same class. This will quantify the opinions for each entity. The main focus of the system is identifying semantic orientation of words in the first phase, using WordNet.

Generally, sentiment analysis is only done for subjective statements, and objective sentences are overlooked. Jalaj S Modha et al. [21] describe, among others, an approach to handle both objective and subjective sentences. Classification techniques are used to classify sentences as opinionated or non-opinionated, and further, opinionated sentences are classified as objective and subjective. Specific grammar rules and semantic orientations are defined for objective analysis. A combination of POS tagging, WordNet and SentiWordNet and a large comprehensive lexicon specifically for objective sentences will allow for sentiment analysis to be performed for objective sentences.

There are several challenges however in sentiment analysis, as described by Bing Liu [16]. Some of them are- object identification, feature extraction and synonyms grouping (difficulty in identifying feature of objects), opinion orientation classification and integration of all the above. G.Vinodhini and R.M.Chandrashekaran [19] also elucidate on these problems. The first is that a certain opinion word can have different orientations in different sentences, and thus is not standardized. Another issue refers to the contradictory nature of the statements that people post. A third issue is the opinion summarization task, which differs from traditional text summarization in that only the features of the product are mined on which the customers have expressed their opinions. Thus, new techniques have to be derived to isolate the relevant opinions.

2.4. *TWITTER MINING*

A potential vast source of unstructured web data are *Twitter feeds*. With proper text mining and NLP techniques, large amounts of useful knowledge can be gathered from tweets. Due to the nature of tweets, in which people freely express their opinions, *opinion mining* can be performed, and its results prove useful to several people to get a sense of the sentiments of the public. Apoorv Agarwal et al. [10] describe a '*tree kernel*' structure to perform sentiment analysis on tweets and classify them as positive, negative or neutral. Polarity of words is established prior to the main analysis using WordNet. This prior polarity is used in creating the tree kernel to break down a sentence into its parts-of-speech. The sentiment of the adjectives will specify the overall sentiment of the tweet. Another method is described by Alexander Pak and Patrick Paroubek [14]. Instead of a tree structure, a *sentiment classifier* is built using a training data set. It is based on the Naive Bayes Classifier, and uses N-gram and POS-tags as features.

With respect to Twitter mining, sentiment analysis combined with *association mining* can lead to more useful, topic-specific analysis. Using association rule mining, the expressed sentiments and opinions in tweets can be matched with the entity/ topic being referred

to. Thus, the sentiments attached to the various mentioned entities can be gauged and summarized. Such specific analysis is more useful in the context of social media mining, than compared to commonplace NLP tools, which usually incorporate generic sentiment analysis only, as explained by Alan Ritter et al. [8].

However, he issues with sentiment mining in twitter are diverse, and arise mainly from the informal language structure of tweets. Albert Bifet and Eibe Frank [17] list out some issues with the same. Some of them are- the difficulty in processing twitter-specific texts, like hashtags and user names, processing tweets of a sarcastic/ironic tone, unbalanced nature of tweet classification etc. For example, a major problem experienced is identifying the context of the tweet about the topic, to determine the actual sentiment expressed by the tweeter. Context identification is not straightforward, since a lot of background information about the tweet and its origins is required before the actual sentiment conveyed can be identified.

3. CONCLUSION

With the technology boom we are experiencing, tremendous amounts of data are being generated from Smartphones, blogs, Social Networking sites etc. Experts estimate that the data generated over the past two years is the same as the amount that was generated from the beginning of time, up until 2012. This exponential increase in the amount of data (termed Big Data) has brought with it a number of issues related to data management and information extraction.

This boom in data can however, be looked upon as a blessing rather than a curse. As the web and its usage continues to grow, so does the opportunity to analyse web data and extract all kinds of useful knowledge from it. Many commercial, educational and scientific applications are increasingly dependent on methodologies to extract information from such data sources. Once obtained and analysed, engineers can make many important predictions and discover fascinating trends that would not have been possible before the analysis.

Keeping this in mind, we have researched methods of obtaining web data, followed by the analysis of the information, along with their limitations and best practices to be followed.

4. ACKNOWLEDGEMENT

The authors would like to acknowledge and thank Technical Education Quality Improvement Program [TEQIP-II], BMS College of Engineering and SPFU [State Project Facilitation Unit], Karnataka for supporting the work reported in this paper.

REFERENCES

1) Subashini S, Mahesh T.R, "Web Mining: Prominent Applications and Future Directions", IRACST - *International Journal of Computer Science and Information Technology & Security* (IJCSITS), ISSN: 2249-9555 Vol. 2, No.4, August 2012, pp. 825-830.

2) Subhendu Kumar Pani, Deepak Mohapatra, Bikram Keshari Ratha, "Integration of Web Mining and Web Crawler: Relevance and state of art", (IJCSE) *International Journal on Computer Science and Engineering Vol. 02*, No. 03, 2010, pp. 772-776.

3) Gautam Pant, Padmini Srinivasan, and FilippoMenczer, "Crawling the Web".

4) Carlos Castillo, Ricardo Baeza-Yates, "Practical Issues of Crawling Large Web Collections".

5) Mahesh T R, Suresh M B, M Vinayababu, "Text Mining: Advancements, Challenges and Future Directions", *International Journal of Reviews in Computing* (IJRIC), ISSN: 2076-3328, 2009-2010, pp. 61-65.

6) Giuseppe Carenini, Raymond T. Ng, Ed Zwart, "Extracting Knowledge from Evaluative Text", *Proceedings of the 3rd International Conference on Knowledge Capture*, ISBN:1-59593-163-5, pp. 11-18

7) Gauri Rao, Chanchal Agarwal, Snehal Chaudhry, Nikita Kulkarni, Dr. S.H. Patil, "Natural Language Processing using Semantic Grammar", (IJCSE*) International Journal on Computer Science and Engineering* Vol. 02, No. 02, 2010, pp. 219-223

8) Alan Ritter, Sam Clark, Mausam and Oren Etzioni, "Named Entity Recognition in Tweets:An Experimental Study".

9) Albert Bifet," Mining Big Data in Real Time", Informatica 37 (2013), pp. 15–20.

10) Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data".

11) Pavalam S M, S V Kashmir Raja , Felix K Akorli and Jawahar M, "A Survey of Web Crawler Algorithms*", IJCSI International Journal of Computer Science Issues*, Vol. 8, Issue 6, No 1, Nov 2011, pp. 309-313.

12) S.S. Dhenakaran, K. Thirugnana Sambanthan, "Web Crawler-An Overview", *International Journal on Computer Science and Communication* (IJCSC), Vol. 02, No. 01, Jan 2011, pp. 265-267.

13) Xindong Wu, Xingquan Zhu , Gong-Qing Wu, Wei Ding, "Data Mining with Big Data".

14) Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining*", Proceedings of the International Conference on Language Resources and Evaluation*,2010, pp. 1320-1326.

15) Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa, "Natural Language Processing (Almost) from Scratch*", Journal of Machine Learning Research,* Aug 2011, pp. 2493-2537.

16) Bing Liu, "Sentiment Analysis: A Multi-Faceted Problem", *IEEE Intelligent Systems*, 2010.

17) Albert Bifet, Eibe Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data".

18) Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena, "Large-Scale Sentiment Analysis for News and Blogs", *International Conference on Web and Social Media (ICWSM)*, 2007.

19) G.Vinodhini, R.M.Chandrashekaran, "Sentiment Analysis and Opinion Mining: A Survey*", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE),* Vol. 2, Issue 6, June 2012, pp. 282-292.

20) Matthew Lease, "Natural Language Processing for Information Retrieval: the time is ripe (again)", *Association for Computing Machinery*, 2007.

21) Jalaj S. Modha, Gayatri S. Pandi, Sandip J. Modha, "Automatic Sentiment Analysis for Unstructured Data*", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE),* Vol 3, Issue 12, Dec 2013, pp. 91-97.