Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/11120
DOI URL: http://dx.doi.org/10.21474/IJAR01/11120

RESEARCH ARTICLE

EXTRACTING INFORMATION FROM NEWS AND SUMMARIZING IT USINGDATA SCIENCE AND ANALYTICS

Sudeet Mhatre[1], Swapnali Patil[2] and Karan Jogi[3]

1. B.E in Computer Engineering, B-101 Moreshwarvihar, Agashi, Virar, Mumbai, India.
2. B.E in Electronics and Telecommunication, R5 Bhaveshwari, New Link Road, Kandivali, Mumbai, India.
3. B.TECH in Electronics and Telecommunication, 6/8 Malad Co-op HsgSoc Ltd., Poddar Road, Malad, Mumbai, India.

Manuscript Info

Manuscript History
Received: 05 April 2020
Final Accepted: 07 May 2020
Published: June 2020

Key words:-
Extracting Information, Data Science, News Analytics

Abstract

Exacting information from news is basically finding and exacting pieces of relevant data from the large unstructured data. Many times information is available,but hidden in the news article and is required to extract to be made available for general people. There are various sources of information available to user and the valuable information is present among the data,but the system which can extract the desired information and present it is beneficial, effective and required to save time and minimize efforts. Even if the users try to be extremely precise while finding data and querying, it can be time-consuming and subject to error and may also miss out valuable information in the process. News analytics is defined as "the measurement of the various qualitative and quantitative attributes of textual information using data science and analytics. Few of these attributes are: sentiment, relevance, and novelty. Expressing news information as numbers allows the manipulation of everyday information in a mathematical and statistical way. The aim of this research paper is to find the information in news article on websites, articles, blogs etc. extract it and anticipate its meaning using data science and analytics and present it. For the purpose of demonstrating the feasibility of the methodology this paper will emphasize on extracting the news regarding stock market and anticipating the meaning of the extracted text.

Introduction:-

Due to Internet, huge volumes of structure and unstructured data are available online. Electronic newspapers, article, blogs etc. are increasingly being accessed and read by users from anywhere, anytime. In India there are about 20 daily newspapers, and many of them make an electronic version available online. Newspapers are a source of authentic and information of the given time. There is a large amount of information available in newspaper articles available on various websites.Web articles contain information about crimes, accidents, politics, cultural events and sports events. There are various sources of information available to user and the valuable information is present among the data but the system which can extract the desired information and present it is beneficial,effective are required to save time and minimize efforts.Even if the users be very specific while finding data and querying it can be tedious and subject to error and may also miss out valuable information in the process. Even though valuable

Corresponding Author:-Sudeet Mhatre
Address:-B.E in Computer Engineering, *B*-101 Moreshwarvihar, agashi, virar, Mumbai ,India.

information is available in human-readable form in online newspapers, articles,blogs etc. software systems that can extract valuable information and present this information are rare. Even if one could manually read through the results and extract relevant information, this process is time consuming and can subject to error. So, this work aims to find information available in online.For example-If a person wants information about Cristiano Ronaldo football skills and background then users will find 100's of article online and have to read them all to gain in valuable insight on the same,which is very time consuming and can result in missing out key piece of information.

In this work, stock market information is extracted and is considered as the domain for the work because stock market information is one of the most important factor for individuals to invest in stock market and buying shares and to access the risk, flow and growth of the market. For example, a new investor want to compare the shares of different stocks and want to invest in stock which have limited risk and better returns and other aspects of it and avoid any unknown errors. Currently, this information is not readily available for users, but these can be obtained from newspaper articles of a market.The system that automatically extracts and presents such information has the potential to be used by the investors of various cities to undertake proactive initiatives to reduce error and predict market.. Additionally, government may use this as government regulates much of the stock market's activity to guard investors and make sure the fair exchange of corporate ownership on the open markets. First, it presents a methodology for extracting stock information from newspaper articles. Second, it pre-process information and create a corpus. Third, it classifies the information and create a metrics to build measures that examine whether the analytics are generating classifications that are statistically significant, economically useful, and stable.And lastly,we summarize the information.

This paper is organized as follows. Section 2 provides a proposed framework. Section 3 presents the methodology employed .Section 4 discusses the case study that were conducted using three different newspaper articles on websites and the results. The conclusions and scope are provided in Section 6.

Proposed framework:

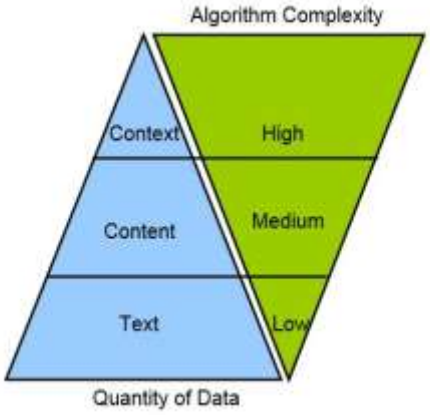

Fig 2.1:- Relation between algorithm and data.

The term "news analytics" consists of the set of techniques, algorithms, and statistics that are used to summarize and classify sources of information.News analytics is a field that related to information retrieval, machine learning, statistical learning, network, and filtering. News analytics consist of three levels: text, content, and context or meaning of that text. The main role of analytics is to convert text into information. This can be done by signing text, classifying it, or summarizing it to reduce it to its main elements. Analytics may even be used to discard irrelevant text, thereby reducing it into information with higher signal content. A second layer is based on content. Content expands the domain of text to images, time, form of text (email, blog, page), format (html, xml, etc.), source, etc. Text becomes enriched with content and asserts quality and veracity that may be exploited in analytics. A third layer of news analytics is based on context. Context refers to relationships between information items. The algorithm has many features, some of which relate directly to text. Other parts of the algorithm relate to content, and the kernel of the algorithm is based on context, News data has three levels: text, content and context. Depending on the layer, algorithms vary in complexity. The simplest algorithms are of those that analyze text alone. Context algorithms, such as the ones applied to network relationships can be quite complex. For example, an Incrementer algorithm is

much simpler, almost naive, in comparison to a detection algorithm. The detection algorithm has far more complicated logic and memory requirements. Complex algorithms produce more structured data. Figure 2.1 depicts this relation. The tension between data and algorithms is moderated by domain specificity, i.e., how much customization is needed to implement the news analytic. High-complexity algorithms are mostly less domain specific than low-complexity ones which are more domain specify.

Methodology:-

This section describes the methodology used for classifying sentences in newspaper articles. Fig 3.1shows the flow of methodology.

Crawling and scraping:

A crawler is a software algorithm that generates a sequence of web pages that can be searched for news content. The word crawler emphasizes that the algorithm begins at some web page, and then moves to branch out to other pages from there, i.e., "crawls from" around the web. The algorithm

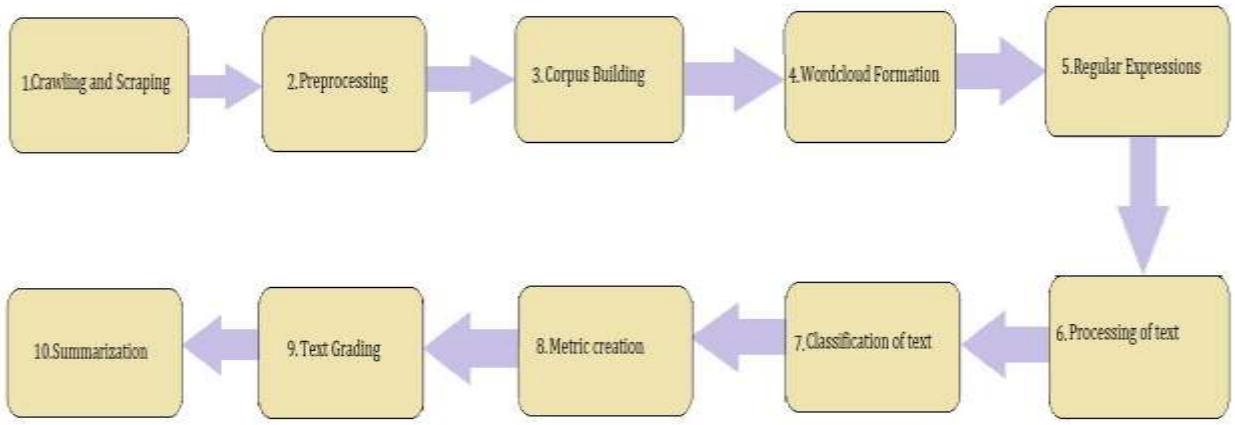

Fig 3.1 Steps followed in the methodology

needs to make intelligent choices from among all the pages it might search into. Commonly used approach is to move to a page that is linked to, i.e., hyper-referenced, from the current page. Crawler explores the tree emerging from any given node, using heuristics to determine relevance of data along any path, and then determines which paths to focus on. A web scraper downloads the content of a web page and may or may not format it for analysis. All programming languages contain modules for web scraping in today's time. These inbuilt functions open a pathway to the web, and then download user-specified URLs. The growing statistical analysis of web text for information contents has led to most statistical packages containing inbuilt web scraping functions.

Preprocessing of text:

Texts from web pages obtained by crawler are usually dirty. Algorithms are needed to clean up before news analytics can be applied. This process of cleaning the text is known as preprocessing. First, there is "HTML Cleanup," to removes all HTML tags from the body of the message. Second, we expand abbreviations to their full form to get clear representation, making the representation of sentences with abbreviated words common in the message. For example, the word "isn't" is replaced and changed to "is not". Third, whenever a negation word appears in a sentence, it usually causes the meaning of the sentence to be the opposite of that without the negation.for example, "the place is not good" means opposite of the sentence "the place is good".

Building corpus:

To create the corpus we will use the tm package. Tm package's main data structure is corpus which consists of collection of text documents.The key benefit of constructing a corpus using the tm package is that it provides you the ability to run text operations on the entire corpus, rather than on just one document at a time which can save plethora of time.

Forming wordcloud:
A word cloud is a collection, or cluster, of words depicted in different sizes in the given corpus. The bigger and bolder the word appears, the more often it is mentioned within a given text and the more important it is.These are ideal ways to pull out the most relational parts of textual data, from blog posts to databases. They can also help business users compare and contrast two different pieces of text to find the wording similarities between the two.

The frequency of a word in a document is defined by the formula as
$f(w,d) = \#w \in d/|d|$
where |d |is the number of words present in the document.

Regular expressions:
A regular expression or rational expression is a collection of characters that is used for finding search patterns. Usually these patterns are used by string searching algorithms for "find" or "find and replace" operations on string of words, or for validating input. Regular expressions are syntax used to modify these strings in an efficient manner. They are complicated but very effective.

Text Processing:
Once search patterns are found in the corpus. Text processing is the process of analyzing text data for getting structured information and manipulationof textual information. This includes extracting bits of information from text, assigning values depending on its content and performing calculations that depend on the textual information from the document. Text processing methods that can be used are-:
1. frequency distribution
2. collocation
3. concordance
4. TF-IDF,

you can make use of all these statistical methods to process and analyze text.

Text Classification:
Text classification classifies text into pre-defined groups depending upon its content, helping to automatically sort and analyze the textual information The main aim is to take a piece of text and assign it to one of a pre-determined set of categories. This classifier is trained on an initial corpus of text that is pre-classified. The classifier is then applied to text to obtain the posterior probabilities of textual categories. The text is then assigned to the category with the highest posterior probability. Some of the most widely used text classification models includes-:
1. Topic analysis
2. Sentiment analysis
3. Bayes classifier
4. Word count classifier
5. Intent detection
6. Language classification.

Metric creation:
Metrics are necessary after developing analytics. Measures must be formed to assess whether or not the analytics that generates classification important statistically, stable and useful in terms of economy. For an analytic to be statistically valid, it should match some conditionsthat states classification is accurate and significant. As the range of analytics grows, range of metrics will also increase. Following are metrics we form -:
1. Confusion metrics
2. Precision and Recall metrics
3. False Positives metrics
4. Sentiment Error metrics
5. Disagreement metrics
6. Correlations metrics
7. Aggregation Performance metrics
8. Phase-Lag Metrics
9. Economic Significance metrics

Text grading:
Text grading is to assess the readability of text formed after metrics are created. Readability states how easy it is to comprehend the text. This can be extremely helpful while summarizing the text and hence are gaining traction.

Summarization:
Statistical methods make it easy for summarizing the text we got from all the processes. The Sentence based summarizer sorts all the sentences in the text in a descending order of overlapping words with other sentences. This sorting of sentences reorders the sentences and arranges the text with the sentence that has most overlap with other sentences first, then the next, and the process goes on until all sentences are arranged.

An article A may have n sentences wi,i = 1,2,...,n, where each Wi is a set of words.

We calculate the pairwise overlap between sentences using the similarity index:

$Jij = J(wi,wj) = |wi \cap wj| \ |si \cup sj| = Jji$.

The size of the intersection between words of two sets of word in a sentences wi and wj is written in ratio and divided by union of two sets and result obtain is the overlap.

The similarity score of each sentence is computed so as to get the row sums.

$Wi = m \sum j=1 Jij$

The obtained row sums are then sorted out and the summary is the first n sentences based on the Wi values. We can then decide how many sentences we want in the summary as per required. Basically by this process we get most précis information according to the Input specified in the first step.

Case Study:
In this case study, we took article about stock market stats of month December 2019 from three different online website articles namely NSE India, Economic times, and Money control. We then manuallycollected all the relevant information regarding stock pricing, impact of government policy change and market scaling etc. which we expect our proposed system to give as output. We then give these inputs to the system and the resulting output is compared with our manually collected data and then we calculate the accuracy of our system in terms of percentage, the more the percentage of accuracy the more precise the system is. Following table shows the resulting accuracy we received after calculation.

Table 4.1:-

Input Articles	Accuracy(in %)
NSE India	88
Economic times	82
Money Control	84

(Data taken from articles of the given source of December 2019)
The above table shows the accuracy in percentage we get after comparing the manual output with the system output.

Conclusions:-

This paper presents a methodology for extracting news sentences from online newspaper articles. It employs crawler algorithm to find all the information pertaining to the required subject. Then the information gather by the crawler is fed to scraper algorithm to eliminate irrelevant piece of information. Then the information is preprocessed to remove tags and expand abbreviation. Next we build the corpus of the data gather so far from different sources. We use the corpus for forming the word clouds. Then we find regular expressions of the text for finding patterns, which is processed and classified and after classifying we build metrics which is graded for readability and lastly we summarize the resulted information to get the relevant important information. The system proposed is evaluated on four newspaper articles from three diff erent stock market websites. It demonstrates that the accuracy of the results obtained for articles varies from 82% to 88%. This system can be used across various domains like sports, crime,

finance etc. to get relevant information to study other domains. This can be very beneficial and effective in the field of information retrieval where information has to be extracted from large amount of data sources.

References:-

1. Adrian Kay (2019) "Extract information from news articles by using Web-scraping and NLP" https://medium.com/@adrianhkay/extract-information-from-news-articles-by-using-web-scraping-and-nlp.
2. Bishop Carmen. (1995). "Neural Networks for Pattern Recognition".
3. Cowie, J. &Lehnert, W. (1996), 'Information extraction', Communications of the ACM,70-92.
4. Das, Sanjiv. (2014). "Text and Context: Language Analytics for Finance",145-230.
5. Economic times (2020) article on stock market statitics.https://economictimes.indiatimes.com/markets/stocks
6. Inés Roldo (2019) "Text processing and classification using analytics" Monkey learn blog https://monkeylearn.com/blog/author/ines/.
7. Money control (2020) article on stock market https://www.moneycontrol.com/stocksmarketsindia/
8. National stock exchange article on stock market(2020) https://www.nseindia.com/market-data/live-equity-market
9. Patil, D.J. (2011). "Building Data Science Teams".
10. RexyArulanandamBastin Tony Roy Savarimuthu Maryam A. Purvis(2014) "Crime Information from Online Newspaper Articles"
11. Sanjivranjandas(2013) book on "Data science: theories, models, algorithms, and analytics",184-221.
12. Seigel, Eric (2013). "Predictive Analytic", New Jersey.