# Differentiating objective and subjective semantics using Text mining and Web Scraping: A Comparative Study

Sachit Misra, Anushka Agarwal

**Abstract**— The active and rapid production of a large amount of digital data and the need to rearrange it into a readable and usable format is not feasible manually. However, using Web Scraping and Automation, a considerable amount of data can be collected, and the analysis of that data can draw reasonable outcomes. After extracting the desired text from the data source, the quantity of output almost makes it impossible to be interpreted without computational support. To conclude a useful result from the data, it has to be statistically identified and later classified based on the main sentiments expressed in that text. A comparative study between two text narratives alongside Sentiment Analysis and Trend Analysis can be used to get a brief insight into the difference of sentiments between common public opinions and well-established facts. However, for it to be implemented on a vast scale, the availability of a large dataset and a proper linguistic approach to analyze the dataset becomes a necessity. This paper accommodates a comparative study of text mining on factual text from news sources and literary text collected from user comments and responses being tested against multiple NLP Libraries. Sentiment analysis is computationally identifying the opinions portrayed in a piece of a text abstract to govern a person's attitude towards a product. Another very prominent approach in text mining is Trend Analysis, which mainly deals with the extraction of consequential keywords and their frequency of occurrence from a substantial volume of unstructured text. Being aware of the current trends around multiple text classification genres gives a brief insight into that particular topic and the present-day public demands.

**Index Terms**— Text mining, Sentiment Analysis, Natural Language Processing, Web Automation, Linguistics



## 1 Introduction

Considering a significant section of the web embraces text, it can be comprehensively classified into two major domains - Factual and Literary. The Factual Text deals with facts, figures, and statistics, which might result from some scientific hypothesis, current affairs, geopolitical interference, or economic policies. Social media is one of the most significant information exchange technologies of the 21st century. People of all ages use social media to post messages, photos and videos about their daily activities. Social media channels, such as Twitter and Facebook, provide very convenient and efficient ways of communicating and sharing information publicly.[1] The literary text comprises public opinions, points of view, or judgment.

With the existence of both domains, it becomes very pertinent to have a significant distinction between the two.

The world wide web has made it feasible for people aside from the mainstream media to state their opinion on any subject that rests on any webpage which has the provision to post comments. This goes hand in hand with the companies whose sustenance depends on what people post about their products. It would be beneficial for these companies as well as sellers to know what the general population thinks of their products as it is vital for the success of their business. Since the internet is a vast place that has billions of gigabytes, it is difficult to fulfil the desire of these businesses to know what people think of their product. It is therefore helpful to employ a data collection bot to perform this task on select websites that might hold some information useful to these businesses.

A pronounced implication of Text Mining is to draw out data-driven decisions, which means using the auto-generated user's everyday data into meaningful hypotheses by studying the common patterns involved with that complex unstructured data. For an individual to conclude from a data, the availability of a sufficiently large data set is a significant challenge. Following the multiple text divisions, both kinds of text are scattered all over the internet.

The most convenient way of scraping data out of a webpage is by using a web crawler that traverses through the page using specific Python libraries and working its way through the page's HTML. Web Scraping refers to how selective information can be concluded from a webpage and later used to produce the desired result. To get started with text extraction from websites, a python library, selenium could be employed to perform this task.

After a convenient amount of data is scraped from the selective web pages, multiple Python libraries can be used to determine the Sentiment and Trend Analysis. These lexicon-based libraries make use of Natural Language Processing to categorize words relating to sentiment as well as the frequency of the expression over a broad span of text. After the data is successfully collected, it is specified under three complete categories- Positive, Negative, and Neutral. In other words, Document-level sentiment classification automates the process of classifying a textual content, which is given on a single topic, as expressing a positive, negative, or neutral sentiment.[2]

In this paper, a comparative study is concluded between the distinctions of text on immensely popular topics in the form of multiple case studies, which are compared against the efficiency of numerous Text processing libraries. To achieve a dataset that would serve as the most accurate representation of general public opinion, we have used Reddit's official website (https://www.reddit.com/), taking all the subreddits belonging to the case study topics into consideration. For factual

text, we have considered using Euronews (https://www.euronews.com/) and scraping all the articles within the range of the same subject.

### 2.1 Related Work

Natural language processing (NLP) has recently gained much attention for representing and analyzing human language computationally.[2] NLP is emerging out as the current Machine Learning trend with its applications and roots getting down to Artificial Intelligence, computational linguistics and computer science. Sentiment and Trend Analysis is just a small subset of its actual application, which is steadily increasing, from Automatic Text Summarization to Chatbot every linguistic related domain is analyzing more efficient algorithms to increase its efficiency. Automatic text analysis become an integral part of many systems, pushing boundaries of research capabilities towards what one can refer to as an artificial intelligence dream - never ending learning from text aiming at mimicking ways of human learning.[3] Automatic text processing is a research field that is currently extremely active. One important task in this field is automatic summarization, which consists of reducing the size of a text while preserving its information content [4]. From classification to deep learning various ML Models are deployed and tested.

### 2.2 Machine learning training classifiers

This approach is used to build a model by feature selection or by learning from a labelled training dataset [5]. It is broadly divided into supervised and unsupervised learning methods. The goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown.[6] Supervised learning algorithms can learn features for classification through data provided in a specific field and optimization. Linear Classifiers, support vector machines, decision trees like random forest and probabilistic classifiers like Naive Bayes Classifier are some ways of implementing machine learning sentiment analysis. Naive Bayes is a classification algorithm which is based on Bayes theorem with strong and naïve independence assumptions. It simplifies learning by assuming that features are independent of given class [7]

In unsupervised learning, previous assumptions and definitions are not provided to the model. The preprocessed data is examined, and the structure of the data is learnt by the algorithm.

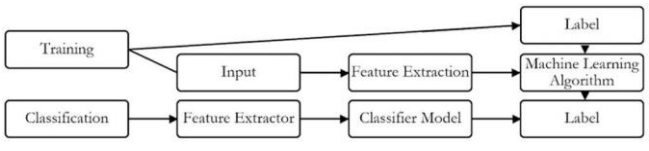



Fig. 1. Process of labeling tokens via training and classification

Implementing machine learning algorithms in sentiment analysis yields accurate results but the time taken to construct pipelines and generate a training dataset can be a drawback for small businesses. Since the sentiment classifier is trained on the labelled data from one field, often it does not work with another field efficiently because of missing information about the other field's usage of words and its weights. To solve this issue, lexicon-based approach is used as it caters to a wider section of usage and implementation.

### 3.1 Related Work

Web scraping is the set of techniques used to automatically get some information from a website instead of manually copying it. The goal of a Web scraper is to look for certain kinds of information, extract, and aggregate it into new Web pages. [8] The fundamental approach of Text mining is data collection, which is executed by using selenium, a python library. One needs a basic understanding of HTML, some knowledge of web-based Structure, and the necessary developer tools following a web Driver specialized with selenium to carry out this extraction process. Selenium uses a web browser to imitate user-governed actions on a webpage for primary navigation. It enables python to easily administer the browser via operating-system-level interactions, which aids in data collection automation. An automated bot is deployed on the websites, and the data is scraped out by following and manipulating the page's HTML. The bot is made to transverse the list of articles or comments present in a website, and then it is processed under Text Mining Libraries.

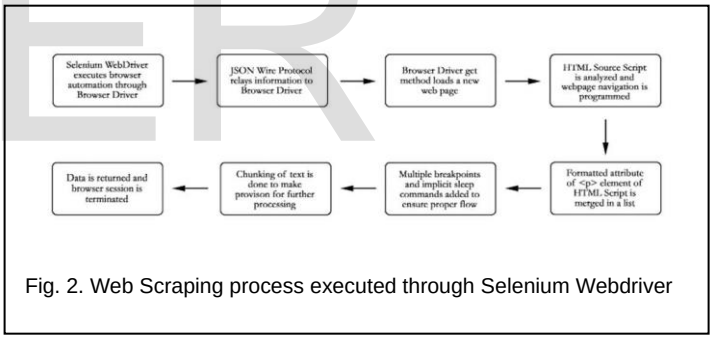



Fig. 2. Web Scraping process executed through Selenium Webdriver

The first step in the evaluation is the Tokenization of a given string. In order to convert the noise in the text from high dimensional data to a low dimensional space, the text is converted to Tokens before transforming them into vectors. The data thus collected contains a lot of recurring irrelevant words that are not at all contextual to the text. By the process of Stop Word Removal, all such words are separated from the text leaving text with significant keywords only, which would later be used for analyzing the Sentiment and Trend Analysis. The remaining data still consist of a combination of letters, words, and punctuations, which often complicates text mining and should be excluded by punctuation removal.

Occasionally, the stem part of the word is subjected to the addition of affixes such as -ed, -size, -s, -de, and miss, decreasing the process's efficiency and hence needs to be lemmatized. With Lemmatization, the given word is normalized and returned to its root form. After being lemmatized the text array is grouped by the process of POS tagging where the tokens are classified on their parts of speech belonging to a spe-

cific language. Depending upon the requirements, the whole set of classified keywords is separated and processed under further procedural algorithms. While undergoing Pos tagging, the words are stratified into two significant divergences, Nouns and Adjectives for Trend and Sentiment Analysis.

### 3.2 Sentiment Analysis

Sentiment analysis is the process of computationally identifying and categorizing the opinions expressed in a piece of text, especially in order to determine the writer's attitude towards a particular topic, product, etc. is positive, negative or neutral [9] Its efficiency is determined by its Sentiment Library, which can uncover sentiments and scores in words or phrases. These Sentiment Analysis libraries are pre-scored data based on adjectives which are scored manually. While working out its way to sentiment Analyze a given text, a predefined algorithm needs to be accurately processed so that it is capable of differentiating the extent to which the text approaches the positive and negative polarity and, hence concluding the results. After the text collection and preprocessing the data.

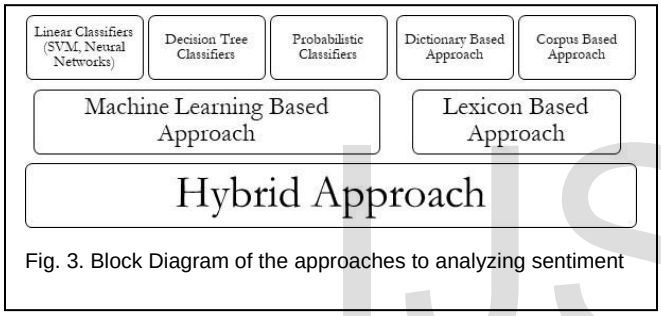



Fig. 3. Block Diagram of the approaches to analyzing sentiment

### 3.3 Simplified Trend Analysis

Forming a net of inter-related words is beneficial as it provides further information on the subjects' sentiment and relations. Empirically, the nouns hold most of the subject matter in chunks of textual information. A rough net of words used to accompany the subject, provided the frequency of the nouns are given, can be drawn. To obtain this frequency, the nouns can be obtained from the POS tagging and arranged in an ascending order. This gives us the nouns most used with the subject.

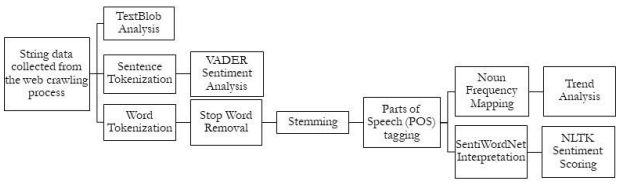
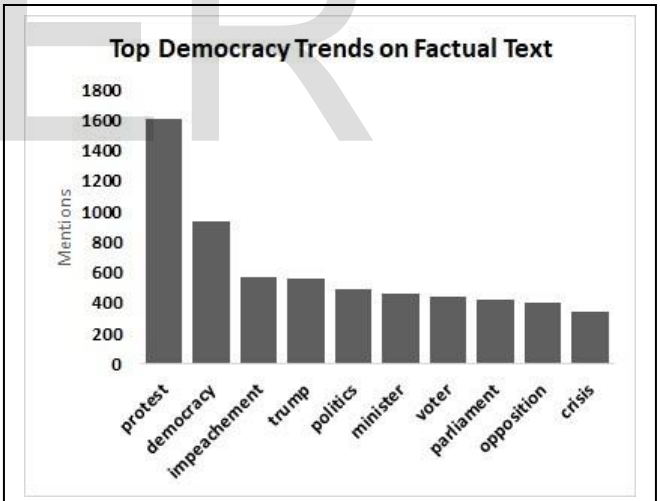


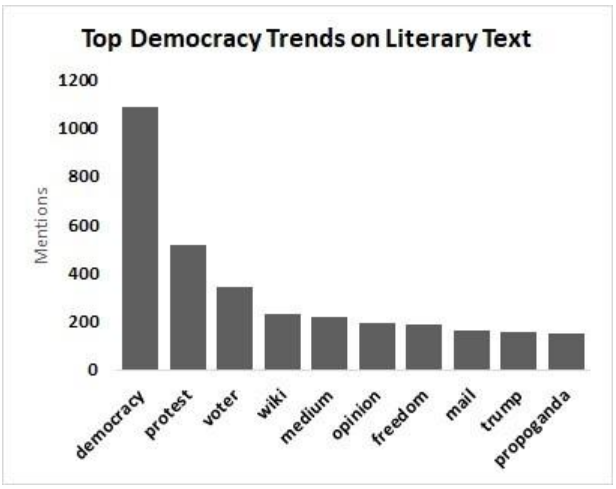

Fig. 4. Preprocessing of string data for different libraries

## 4 Result And Discussion

For the demonstration of this method some topics of controversy were chosen to find out the polarities and the trends in factual and literary text. Extracted out of the aforementioned websites, the data was collected via a script that included a python method which carried out the entire extraction process. This process took about 4 to 5 hours to gather data of 1000-1100 from the news website (https://www.euronews.com) with an average of 400000 words and 2500000 characters and about the same time to collect 15000-35000 comments from Reddit (https://www.reddit.com/) on the same topic. An additional computational processing time of a few minutes was taken by the python libraries for the analyses. On an average the POS density of nouns in the text, defined by the number of nouns divided by the total number of words, was nearly 39% for factual text and 27% for literary text. For POS density of adjectives, the figures were 5% for factual text and 23% for literary text. This suggests that factual text utilizes more nouns and a smaller number of adjectives as opposed to literary text which employs more adjectives and fewer instances of nouns. For trend analysis the frequency of nouns was mapped to form a graph displaying ten of the highest number of mentions. This metric indicates the prevalence of there being a subject that is being referred to which must mean that it is either directly or indirectly related to the topic of discussion and hence ranks up on the relevance as the number of mentions increase. The noun frequency mapping labeled trend analysis is especially useful to form a net of interrelated words which add to the dimension of the subject.




(Continued to next page)



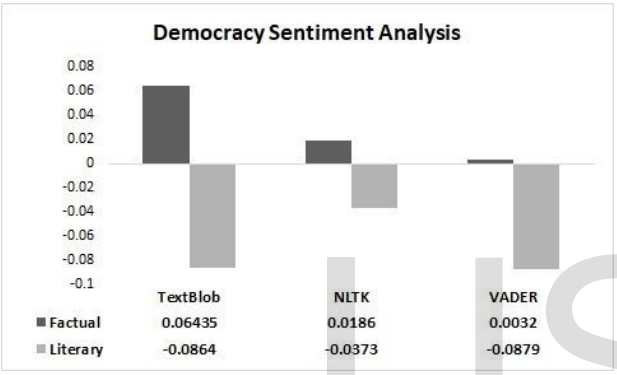
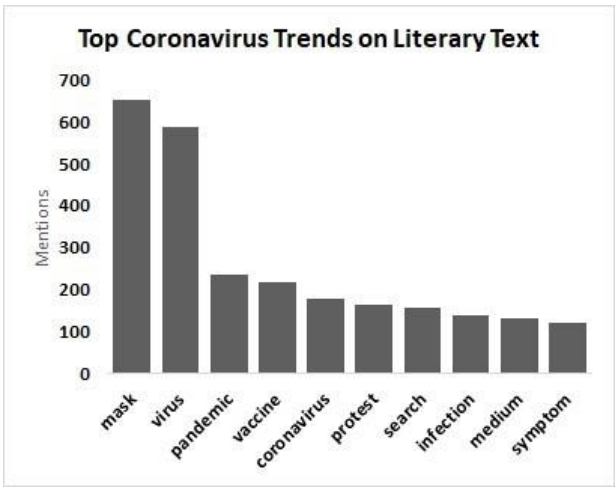




Fig. 5. A total of 1026 articles were extracted on 24th July 2020 from the news website and 21852 comments on the same date from Reddit on the topic of “democracy”. For this topic, the sentiments seem to be strongly polarized with the sentiment analysis libraries showing an overall positive sentiment for literary text and a negative sentiment for factual text. The data suggests that sentiments of the public are not in agreement with the articles provided by the television news network - Euronews. In the trend analysis pertaining to this theme, the word “protest” had the highest association with the topic precluding the name of the theme itself.



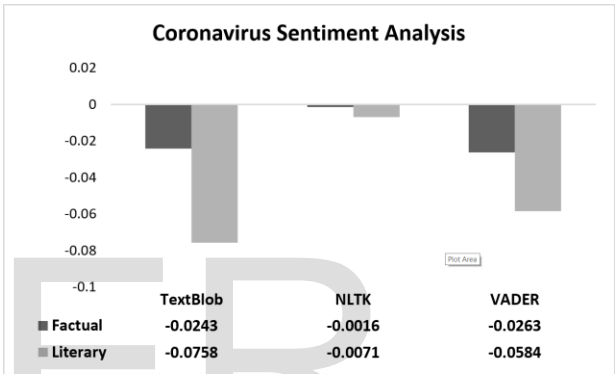
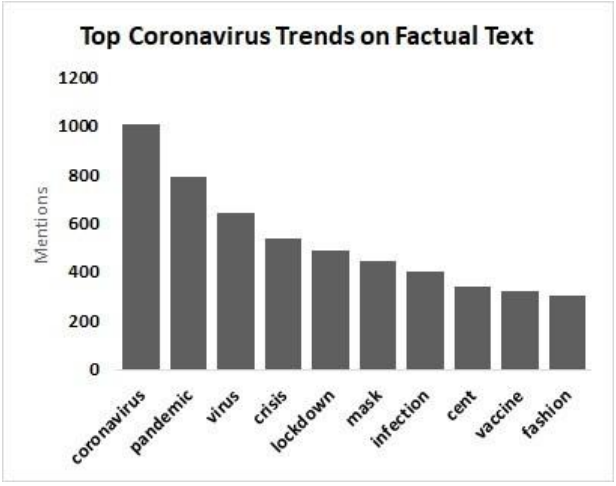
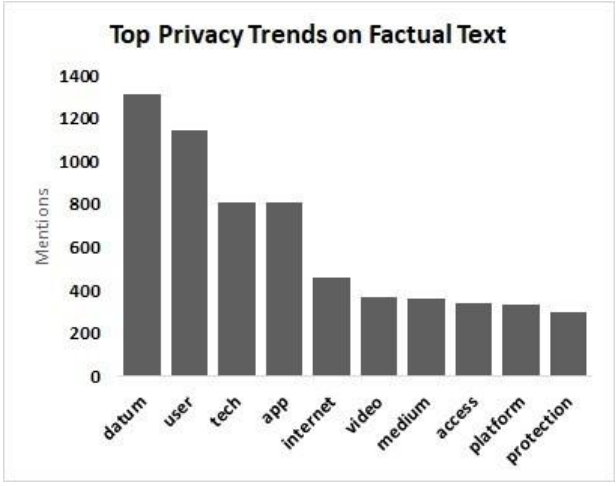



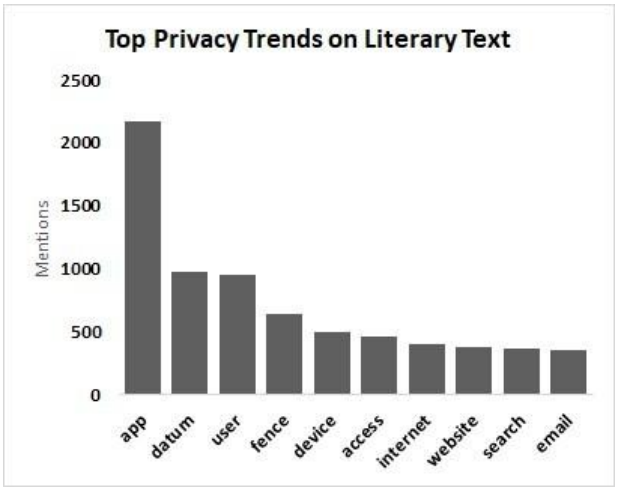

Fig. 5. A total of 1014 articles were extracted on 29th July 2020 from the news website and 17521 comments on the same date from Reddit on the topic of “coronavirus”. The sentiments regarding this theme appear to unify all the sentiment analysis libraries to display a strictly negative outcome. The figures indicate that sentiments of the public are in line with the news. In the trend analysis concerning this theme, the word “pandemic” had the highest correlation with the topic barring the name of the theme itself.




(Continued to next column)




(Continued to next page)



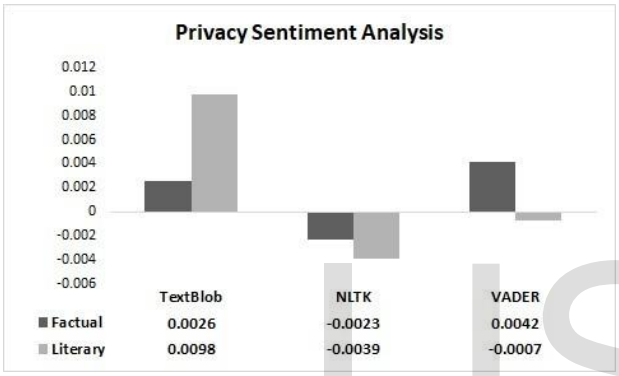
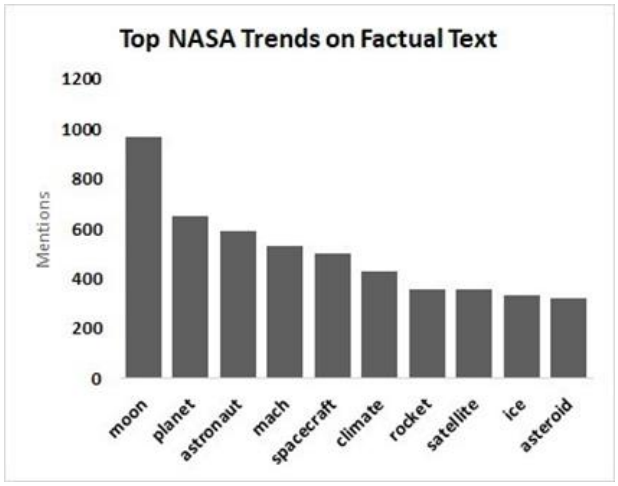
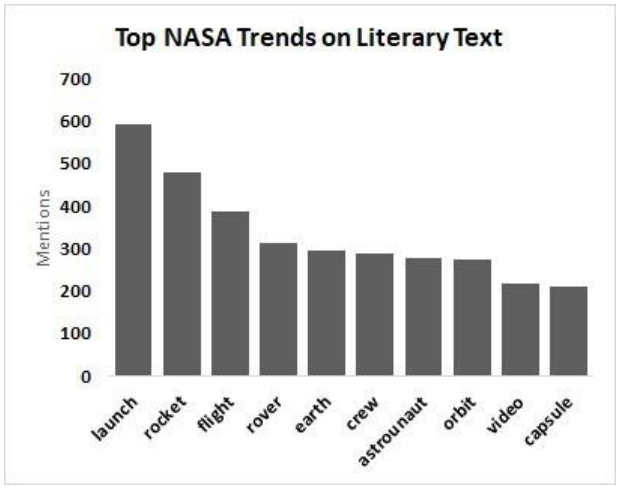




Fig. 5. A total of 1054 articles were extracted on 1st August 2020 from the news website and 34298 comments on the same date from Reddit on the topic of "privacy". The sentiments concerning this theme appear to align all the sentiment analysis libraries to display an overall neutral result. In the trend analysis with reference to this theme, the word "app" had the highest correlation with the topic of discussion.




(Continued to next column)



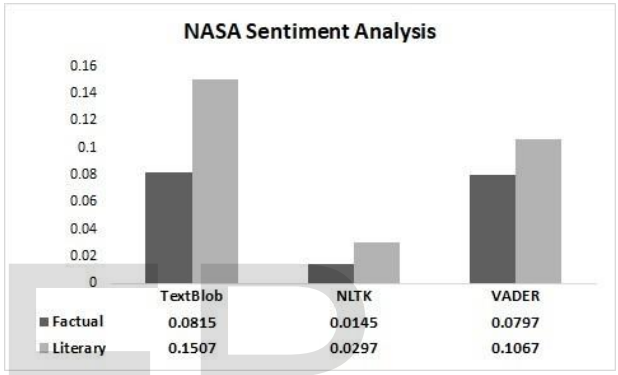




Fig. 5. A total of 1283 articles were extracted on 3rd August 2020 from the news website and 17819 comments on the same date from Reddit on the topic of "NASA". The sentiments concerning this theme appear to align all the sentiment analysis libraries to display a purely positive result. The figures point out that sentiments of the public concur with the news. In the trend analysis with reference to this theme, the words "Launch" and "Moon" had the highest correlation with the topic.

## 5 Conclusion

The research in this paper used Lexicon based libraries which solemnly work with the text weightage associated with the polarity set during the manual execution of the words and phrases. While using such Lexicon based approaches the need to generate massive and complex training dataset can be eliminated to a large extent. Apart from that these approaches are more beginner friendly and can be very easily modified. We infer web scraping can increase access to any textual data available on the internet without depending on, dramatically increasing dataset, decreasing the time spent on data collection phase, increasing access to researchers on the vast amount of data available on the internet, and improving the interdisciplinary application of immense research literature on different fields of study. After the data collection phase, pre-processing of the textual data to strip the text of noise while scraping the HTML source script can be performed by a plethora of libraries. Furthermore, investigation on the polarities of three sentiment analysis libraries specializing in different forms of tex-

tual data was carried out. For the trend-based approach, a simplified trend analysis which revealed the most used terms with that particular topic was performed. We tentatively conclude that sentiment analysis on factual text remains to be presented in a neutral manner whereas literary text involves more emotion. In the future, different languages can be analyzed to provide the aforementioned statistics.

## REFERENCES

[1] Jurek, A., Mulvenna, M.D. & Bi, Y. Improved lexicon-based sentiment analysis for social media analytics. Secur Inform 4, 9 (2015).

[2] Vala Ali Rohani and Shahid Shayaa, "Utilizing Machine Learning in Sentiment Analysis: SentiRobo Approach "(2015)

[3] Mladenić, Dunja & Grobelnik, Marko. (2013). Automatic text analysis by artificial intelligence. Informatica (Slovenia). 37. 27-33.

[4] Khurana, Diksha & Koli, Aditya & Khatter, Kiran & Singh, Sukhdev. (2017). Natural Language Processing: State of The Art, Current Trends and Challenges.

[5] Luhn, H.P. (1958). The Automatic Creation of Literature Abstracts. IBM J. Res. Dev., 2, 159-165.

[6] Kotsiantis, Sotiris. (2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica (Ljubljana). 31.

[7] Fragos, Kostas & Belsis, Petros & Skourlas, Christos. (2014). Combining Probabilistic Classifiers for Text Classification. Procedia - Social and Behavioral Sciences. 147. 10.1016/j.sbspro.2014.07.098.

[8] M. Hu and B. Liu, "Opinion Extraction and Summarization on the Web", pp. 1621–1624.

[9] D. PRATIBA, A. M.S., A. DUA, G. K. SHANBHAG, N. BHANDARI and U. SINGH, "Web Scraping And Data Acquisition Using Google Scholar," 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2018, pp. 277-281, doi: 10.1109/CSITSS.2018.8768777.