IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SENTIMENT ANALYSIS TECHNIQUE AND NEUTROSOPHIC THORY FOR MINING AND RANKING BIG DATA FROM ONLINE REVIEW

Ms.C.Manju Priya, II ME CSE - Hindusthan College of Engineering and Technology
Mr.Jayaraj, M.E , Assistant Professor, CSE - Hindusthan College of Engineering and Technology
Dr.S.Uma, Research coordinator, CSE- Hindusthan College of Engineering and Technology

Abstract

A huge amount of data is being generated everyday through different transactions in industries, social networking, communication systems etc. Big data is a term that represents vast volumes of high speed, complex and variable data that require advanced procedures and technologies to enable the capture, storage, management, and analysis of the data. Big data analysis is the capacity of representing useful information from these large datasets. Due to characteristics like volume, veracity, and velocity, big data analysis is becoming one of the most challenging research problems. Semantic analysis is method to better understand the implied or practical meaning of the input dataset. It is mostly applied with ontology to analyze content mainly in web resources. This field of research combines text analysis and Semantic Web technologies. The use semantic knowledge is to aid sentiment analysis of queries like emotion mining, popularity analysis, recommendation systems, user profiling, etc. A new method has been proposed to extract semantic relationships between different data attributes of big data which can be applied to a decision system. The proposed model associates with genetic algorithm and association rule mining in Map Reduce platform to find out features from big data that helps users to make better decisions. The task for finding frequent item sets from big data is very difficult as well as huge time consuming.

INTRODUCTION

Intense growth in the use of internet-based technologies and activities viz., social media communications, blogging, e-commerce, and surveillance, has amassed huge volumes of data. The challenge before the research community is to address the unstructured, semi-structured, and structured data. Big data analysis is one that addresses officiously with huge possibilities of insights. A combination of data mining, text mining, web mining and natural language processing techniques are used in various real-life applications for big data to be analyzed. The volume of data is increasing from multiple sources such as Facebook, Twitter, Amazon, YouTube, etc. The enormous amount of data associated with customer opinions/reviews are quite challenging to analyze, and it needs innovative approaches to get a generalized opinionated summary.

In many web facilities like news reports, e-commerce websites, social media networks, blogs and forums help to express opinions. They can be utilized to understand the opinions of the general public and consumers on social events, political movements, company strategies, marketing campaigns, product preferences and monitoring reputations [Sal, 11]. This research work has mainly focused on the analysis of the state-of-mind (opinion/sentiment) of people on social media data.

Big Data (BD) is an immense part to play with new technology trends in this world. Mainly, BD is stretching unstructured data, which is very difficult to process and analyze. The rapid growth of unstructured data like BD faces countless issues and challenges which are implied in academics, industries and organizations. Especially, BD

challenges are included for analyzing data, sharing the information, storage, curation, visualization and query updating information by dramatically changing the way to enhance and transform the data.

In this world, big data is pushing the boundaries on the scope of analytics and changing dramatically, constantly evolving to redefine it day by day. Big data offers numerous challenges and commercial opportunities for initiative today. Presents the evolution of big data based on the problems raised in the 3V's Volume, Velocity and Variety.

The big data is referred as a dataset having a huge growth. The BD issues and challenges are very complex, by the usage of traditional database management tools. The issues can be related to data capture, storage, search, sharing, analytics and visualization. The most widely accepted explanations and descriptions are specified as occasionally as "big data problem". In 2001, Laney described three characteristics of data management challenges, which addresses volume, velocity, and variety which is frequently documented in the scientific literature Depicts the characteristics of big data. Three V's are detailed as follows:

- Volume (size of data): the massive amount of data generated each second by the public, consumers and organizations. The volume is fixed to grow exponentially through the web. The data scientists are required to rethink about data storage and processing models, to develop the tools needed to analyze the massive amount of data.
- Velocity (streaming data): the speed of data transformation and usage is increasing in correlation with the volume of data, which accelerates storage, processing and analysis. In simple term, it is called as data in motion. The streaming data can be analyzed using batch processing and stream processing. The batch processing will work on historical data, and stream processing analyzes the data in real-time as it is generated.
- Variety (data type): the source of data refers to different data formats. It is also called as a mixed form of data. The data can come from various sources and take on many ways.

"Big Data Analytics (BDA) is the process of examining large data sets containing variety of data types i.e., big data to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information". The big data analytics are helpful to the organizations to take facility.

The following four types of big data analytics need to be considered when debating how the best leverage information within the business.

- Descriptive (what happened)
- Diagnostic (why it happened)
- Predictive (what will happen)
- Perspective (what action to take)

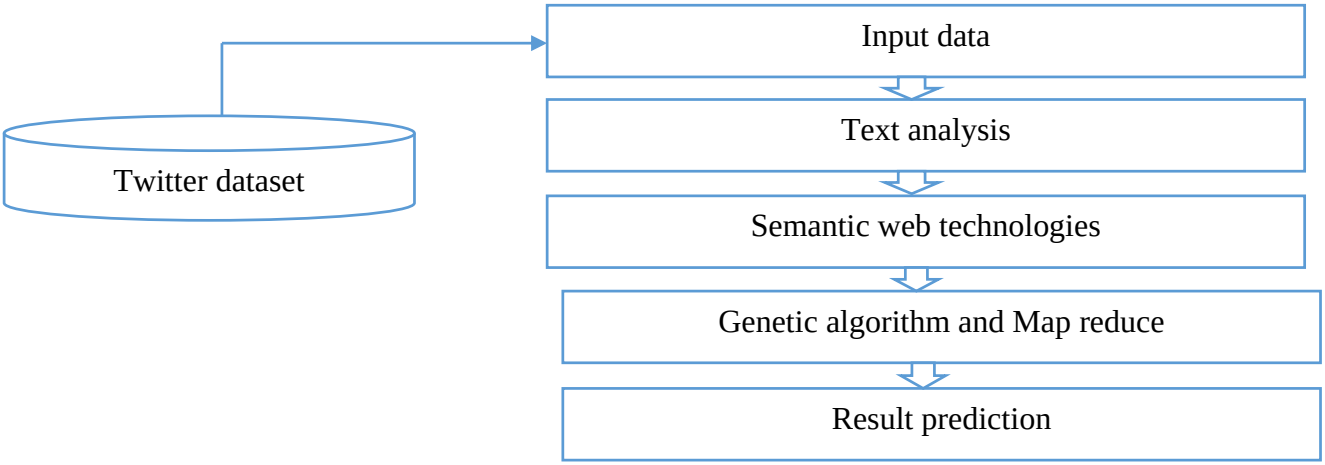
"Sentiment analysis, also called as opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. The previous paints as a problem. There are also many names and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc." [Bin, 12].

Sentiment Analysis (SA) is a computational study of people opinions, sentiments, emotions and attitudes, which are expressed in texts towards an entity. SA facilitates to accomplish numerous goals, which includes locating public mindset, concerning the political occasion, marketplace intelligence, the extent of customer pride and film sales prediction. Obviously, due to the growth of interest in e-commerce, the user can post their comments, opinions and reviews on the social media websites / social media networks. The opinionated data is a rich source to express and analyze the sentiments. The customers / user's opinions are improving the quality and standard of the products, while the reviews are posted by the customer in the e-commerce website/social media networks. For example, the reviews given on e-commerce sites such as Snapdeal, Flipkart, eBay, Amazon, IMDb and epinions.com can affect the customer's decision in buying products and subscribing services.

Online service provides a platform to share the ideas and to cheer people for group discussions with open views. The group discussion gives a better way to get feedback and quick response from people on different issues and entities (in the form of textual posts, news, images and videos). Therefore, the entities can be utilized to analyze peoples' opinions about understanding the consumer's attitude and market trends.

In trust-based security schemes, each node collects two major types of information about other nodes: first-hand information (based on self-observations) and second-hand information (based on the other node observations). In literature, efforts have been made to minimize the bootstrapping time and to increase the detection rate by using second-hand information to evaluate the trustworthiness of the nodes. However, the aforementioned schemes

still suffer from data sparsity problem. In trust-based security schemes, data sparsity is a situation where lack of information or insufficient interaction experience makes it difficult to evaluate the node's trust, especially in the early time of network establishment.



SYSTEM STUDY

Existing System

To analyses the behavior of news required maximum resources. To analyze the fake news, we required manpower to deep down into it and check the authentication of news. We must check all possible connection with news manually. It is time consuming and costly approach.

Disadvantages

- Time Consuming Process
- Manpower Required.
- Deep Knowledge required.
- Cost driven approach.

Proposed system

to detect bots from real-time data by considering the user-based features, content based features and the sentiment score of each user and analyzing how efficiently sentiment analysis score results can be used in detection of bots by using fake prediction algorithm. The proposed system is more effective than the existing one. This is because we will be able to know how the statistics determined from the representation of the result can have an impact in a particular field as well as influence of negativity spread by rumors.

Advantages

- Understand your customers.
- Measuring your marketing campaign.
- Check out the brand concept.
- Find industry leaders and influence.
- Time complexity reduced

MODULES DESCRIPTION

Text analysis

- Textual Information retrieval techniques mainly focus on processing, searching or analyzing the factual data present.
- Facts have an objective component but, there are some other textual contents which express subjective characteristics.
- These contents are mainly opinions, sentiments, appraisals, attitudes, and emotions, which form the core of Sentiment Analysis (SA).

Semantic analysis

- Semantic analytics, also termed semantic relatedness, is the use of ontologies to analyze content in web resources.
- Semantics helps structuring the plain natural language text with formal representation.
- The current system we are developing performs sentiment analysis by hybridizing natural language processing techniques with Semantic Web technologies.

Genetic algorithm and Map reduce

- Genetic algorithm and Map reduce is based on the structuring process through the evolutionary function.

This algorithm uses genetic manipulation, including reproduction, crossover, and mutation, to derive the solution of different iterations.

GA and Map reduce belongs to search techniques that can efficiently evolve the optimal solution in the reduced space.

SYSTEM DESIGN

Introduction

The System Design Document describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human-machine interfaces, detailed design, processing logic, and external interfaces.

Executive summary of the project

This section provides a description of the project from a management perspective and an overview of the framework within which the conceptual system design was prepared. If appropriate, include the information discussed in the subsequent sections in the summary.

Process of System overview

This section describes the system in narrative form using non-technical terms. It should provide a high-level system architecture diagram showing a subsystem breakout of the system, if applicable. The high-level system architecture or subsystem diagrams should, if applicable, show interfaces to external systems. Supply a high-level context diagram for the system and subsystems, if applicable. Refer to the requirements trace ability matrix (RTM) in the Functional Requirements Document (FRD), to identify the allocation of the functional requirements into this design document.

Constraints on the design

This section describes any constraints in the system design (reference any trade-off analyses conducted such, as resource use versus productivity, or conflicts with other systems) and includes any assumptions made by the project team in developing the system design.

Planning for software design

A software module is the lowest level of design granularity in the system. Depending on the software development approach, there may be one or more modules per system. This section should provide enough detailed information about logic and data necessary to completely write source code for all modules in the system (and/or integrate COTS software programs).

If there are many modules or if the documentation is extensive, place it in an appendix or reference a separate document. Add additional diagrams and information, if necessary, to describe each module, its functionality, and its hierarchy. Industry-standard module specification practices should be followed. Include the following information in the detailed module designs:

A narrative description of each module, its function(s), the conditions under which it is used (called or scheduled for execution), its overall processing, logic, interfaces to other modules, interfaces to external systems, security requirements, etc.; explain any algorithms used by the module in detail

For COTS packages, specify any call routines or bridging programs to integrate the package with the system and/or other COTS packages (for example, Dynamic Link Libraries) Data elements, record structures, and file structures associated with module input and output

Graphical representation of the module processing, logic, flow of control, and algorithms, using an accepted diagramming approach (for example, structure charts, action diagrams, flowcharts, etc.)

Data entry and data output graphics; define or reference associated data elements; if the project is large and complex or if the detailed module designs will be incorporated into a separate document, then it may be appropriate to repeat the screen information in this section.

Process of Input design

Input Screen must be design in such a way to give an easy navigation throughout the screen without the violation of the input validation. Input design is the process of converting the user-originated data into a computer-based format. Inaccurate input data are the most common cause of error in data processing. The goal of an input data are collected and organized into a group and error free. Input data are collected and organized into a group of similar data. Once identified, appropriated input media are selected for processing. The design was done with six major objectives in mind

- Effectiveness
- Accuracy
- Ease of Use
- Consistency
- Simplicity
- Attractiveness

The fundamental goal of planning input centers on:

- Controlling the amount of input required
- Avoiding delayed response
- Controlling errors
- Keeping process simple
- Avoiding errors
- Delivering savvy technique for input.
- Accomplishing the most elevated conceivable degree of exactness.
- Guarantee that the information is satisfactory to and perceived by the staff.

Enter the plan objective is to make the information section simple and legitimate as conceivable from errors and opportunity. In entering the information passage, the administrator has to realize each field's space, the field dispersion of the request, and source documents should coordinate. The processor breaks down the information required at that point, and it is acknowledged or dismissed.

The main objective of designing input focuses on

- Controlling the amount of input required
- Avoiding delayed response
- Controlling errors
- Keeping process simple
- Avoiding errors
- Producing cost effective method of input.
- Achieving highest possible level of accuracy.

Ensure that the input is acceptable to and understood by the staff.

The goal of designing input data is to make entry easy, logical and free from errors as possible. The entering data entry operators need to know the allocated space for each field, field sequence and which must match with that in the source document. The processor analyzes the input required. It is then accepted or rejected.

Output Design

The normal procedure in developing a system is to design the output in detail first and then move back to the input. The output will be in the form of views and reports. The output from the system is required to communicate the result of processing to the users. They are also used as the permanent copy for later verifications.

Output Design consideration

The purpose of outputs has been understood and the efficiency of information contained should be analyzed and confirmed. Then the output have been defined in terms of

- Name of the Output
- Content
- Format
- Frequency

Outputs

This section describes of the system output design relative to the user/operator; show a mapping to the high-level data flows described in Section. System outputs include reports, data display screens and GUIs, query results, etc. The output files are described in Section 3 and may be referenced in this section. The following should be provided, if appropriate:

- Identification of codes and names for reports and data display screens
- Description of report and screen contents (provide a graphic representation of each layout and define all data elements associated with the layout or reference the data dictionary)
- Description of the purpose of the output, including identification of the primary users
- Report distribution requirements, if any (include frequency for periodic reports)
- Description of any access restrictions or security considerations

Code Design

A design pattern is a standardized solution to a software design issue or problem which is encountered daily in real-world application development. A pattern focuses on class design and object interaction. Knowledge of design patterns not only prevents having to re-invent the wheel, it allows developers to discuss their work at a higher level of abstraction.

Design patterns have been the bane of my programming existence. I have trouble learning and remembering them. On the one hand, I feel like I have always been following such patterns throughout my career—even before object-oriented languages. On the other hand, I haven't been able to get a good enough handle on patterns and the terminology to be able chat freely about them with my colleagues

Database design

The database design involves creation of tables that are represented in physical database as stored medical information. They have their own existence. Each table constitute of rows and columns where each row can be viewed as record that consist of related information and column can be viewed as field of data of same type.

Framework for system Design

The design effort transforms the detailed, defined requirements into complete, detailed specifications that direct development and testing. Design decisions detail how the system will meet the defined functional, physical, interface, security, and data requirements. At the end of the design process the design is base lined.

The general system characteristics are defined during design. The operating system is established and the automated system packaged into major design subsystems. Inputs and outputs of each subsystem are defined, interfaces to external systems are designed, and administrative activities are established. Security and auditing needs are also addressed.

A more detailed structure of the system is then created based on the subsystems identified by the general characteristics. Each subsystem is partitioned into one or more design units, or modules. The process is described in a structure chart, flowchart, action diagram, pseudo code, or other acceptable format for each design unit, or module. Detailed logic specifications are written for each
module described and data usage is physically defined to the elemental level. Functions requiring user input and approval are completed in this activity.

Throughout the design phase there are a series of check point and review processes. The design is reviewed to verify that it has the following characteristics:

- Is directly traceable to the requirements.
- Describes how the capabilities defined by the requirements will be implemented.
- The SDD includes
- User, human/computer interface design
- System architecture
- Detailed system design
- Data base design including a physical data model and data dictionary.

SYSTEM TESTING

Before applying method to design effective test cases, a software engineer must understand the basic principles that guide software testing. Davis (DAV95) suggests a set of testing principles which have been adapted for use in this book.

- All tests should be traceable to customer requirements.
- Test should be planned long before testing begins.
- Tests pare to principle applets to software testing. Testing should begin “in the small” and progress towards testing “in the page”.
- Exhaustive testing is not possible.

TESTING STEPS

- Unit testing
- Integration Testing
- Whitebox Testing
- Acceptance testing
- Alpha Testing
- Beta Testing
- Blackbox Testing

UNIT TESTING

Unit testing focuses on verification errors on the smallest unit of software design-the module. Using the procedural design description as a guide, important control paths are tested to uncover errors within the boundary of the module.

The module interface is tested to ensure that the information properly flows into and out of the program unit under test. Boundaries conditions are tested to ensure that the module operates properly at the boundaries established to limit of restrict processing.

Integration test

Integration testing is a systematic technique for constructing the program structure while conducting test to uncover errors associated with interfacing. The objective is to take unit tested modules and build a program structure that has been dictated by design.

Whitebox testing

White box testing is some time is called glass box testing, is a test case design that uses a control structure of the procedural design to drive the test cases. Using white-box testing methods, the software engineer can drive test cases that

- Guarantee that logical decisions are on the true and false sides
- Exercise all logical decisions are on the true and false sides

- Execute all loops at their boundaries and within their operational bounds
- Exercise internal data structure to assure the validity

Acceptance testing

Finally, when the software is completely built, a series of acceptance tests are conducted to enable the client to validate all requirements. The user conducts these tests rather than the system developer, which can range from informal test drive to a planned and systematical executed series of tests. These acceptance tests are conducted over a period of weeks or months, there by uncovering cumulative errors that might degrade the system order time. In this process alpha testing and beta testing are used to uncover the errors that only the end user seems able to find.

Black box testing is not an alternative for white box testing techniques. Rather, it is a complementary approach that is likely to uncover different class of errors. Black box testing attempts to find errors in the following categories:

- Interface errors.
- Performances in data structures or external database access.
- Performance errors.
- Initialization and termination errors.
- Incorrect or missing functions.

APPENDIX 2 Screen shots

	Product Name	Brand Name	Price
0	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99
1	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99
2	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99
3	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99
4	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99

	Rating	Reviews	Label
0	5	I feel so LUCKY to have found this used (phone...	2
1	4	nice phone, nice up grade from my pantach revu...	2
2	5	Very pleased	2
3	4	It works good but it goes slow sometimes but i...	2
4	4	Great phone to replace my lost phone. The only...	2

Figure A.1.1 List Product name and reviews.

Alpha testing

The customer conducts the alpha test at the developer's site. The client notes the errors and usage problems and gives report to the developer. Alpha tests are conducted in a control environment.

Black Box Testing

Black box testing focuses on the functional requirements of the software. That is black box testing enables the software engineer to drive a set of input conditions that will fully exercise the requirements for a program.

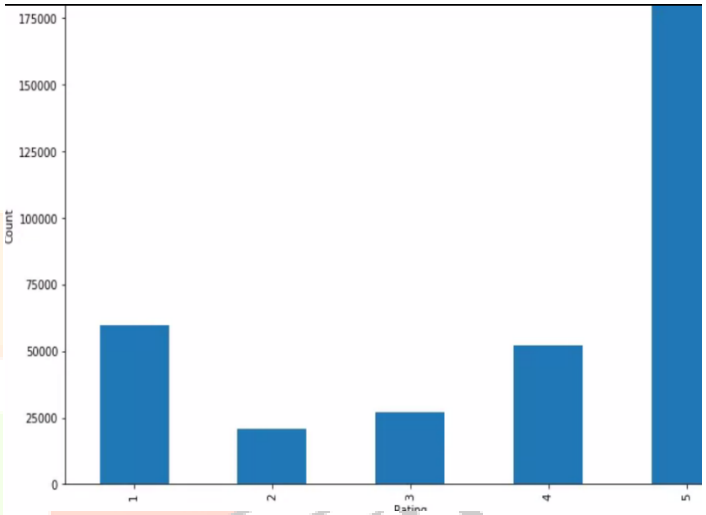


Figure A.1.2 Distribution of product rating

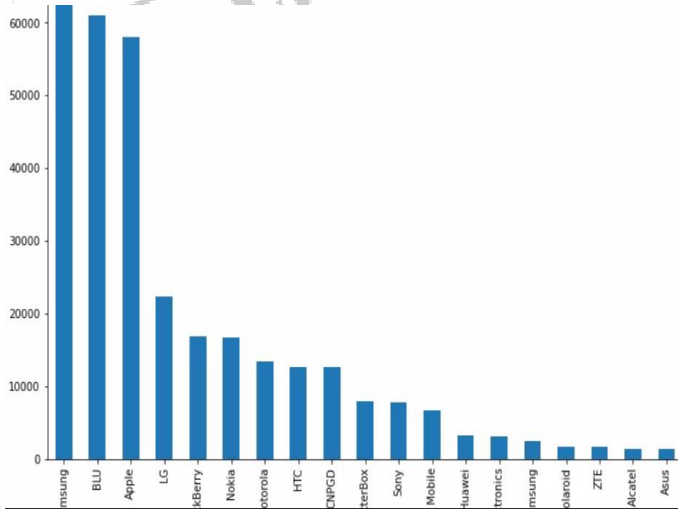


Figure A.1.3 Number of reviews top 20 products

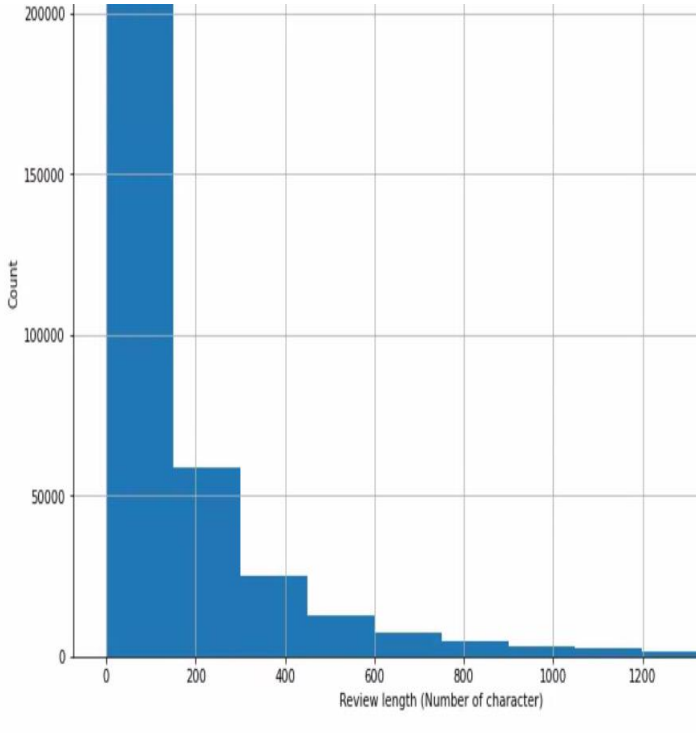


Figure A.1.4 Distribution of Review length

```
Show some feature names :
 ['aa', 'aerial', 'andcamera', 'ascetics', 'baggies', 'bi
o', 'crunched', 'deficient', 'diffcult', 'dong', 'electin
edit', 'hhi', 'ilemming', 'instagraming', 'iwhatever', 'l
celooking', 'office', 'outreach', 'percentbecause', 'pla'
'semipro', 'simscroll', 'solidarity', 'starr', 'superglue
ncmms', 'withxxex', 'yupp']
```

Figure A.1.5 Some features list

Conclusion

Black box testing focuses on the functional requirements of the software. That is black box testing enables the software engineer to drive a set of input conditions that will fully exercise the requirements for a program.

Genetic algorithm and map reduce are used popularly for sentiment analysis. They are majorly used in this field for optimization and search-related problems. Sentiment analysis is the process of analyzing the sentiments or views in a given piece of text. It belongs to the Natural Language Processing domain. The main output of sentiment web based independent classifier for text valence and optimization method based on genetic programming. To make the optimization feasible together with big data approach we have proposed GA operators, which significantly accelerate conversion to the accurate solutions.

Reference

1. I. Awajan, M. Mohamad and A. Al-Quran, "Sentiment Analysis Technique and Neutrosophic Set Theory for Mining and Ranking Big Data From Online Reviews," in IEEE Access, vol. 9, pp. 47338-47353, 2021, doi: 10.1109/ACCESS.2021.3067844.

2. Q. Peng, L. You, Q. Lu and X. Li, "Mining Review Unit Model for Online Review Analysis," in IEEE Access, vol. 8, pp. 196826-196834, 2020, doi: 10.1109/ACCESS.2020.3033820.

3. L. Yang, Y. Li, J. Wang and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," in IEEE Access, vol. 8, pp. 23522-23530, 2020, doi: 10.1109/ACCESS.2020.2969854.

4. Z. Li, R. Li and G. Jin, "Sentiment Analysis of Danmaku Videos Based on Naïve Bayes and Sentiment Dictionary," in IEEE Access, vol. 8, pp. 75073-75084, 2020, doi: 10.1109/ACCESS.2020.2986582.

5. Y. Zhou and S. Yang, "Roles of Review Numerical and Textual Characteristics on Review Helpfulness Across Three Different Types of Reviews," in IEEE Access, vol. 7, pp. 27769-27780, 2019, doi: 10.1109/ACCESS.2019.2901472.

6. S. Hu, A. Kumar, F. Al-Turjman, S. Gupta, S. Seth and Shubham, "Reviewer Credibility and Sentiment Analysis Based User Profile Modelling for Online Product Recommendation," in IEEE Access, vol. 8, pp. 26172-26189, 2020, doi: 10.1109/ACCESS.2020.2971087.

7. S. Ali, G. Wang and S. Riaz, "Aspect Based Sentiment Analysis of Ridesharing Platform Reviews for Kansei Engineering," in IEEE Access, vol. 8, pp. 173186-173196, 2020, doi: 10.1109/ACCESS.2020.3025823.

8. Z. Kastrati, A. S. Imran and A. Kurti, "Weakly Supervised Framework for Aspect-Based Sentiment Analysis on Students' Reviews of MOOCs," in IEEE Access, vol. 8, pp. 106799-106810, 2020, doi: 10.1109/ACCESS.2020.3000739.

9. S. Zhang and H. Zhong, "Mining Users Trust From E-Commerce Reviews Based on Sentiment Similarity Analysis," in IEEE Access, vol. 7, pp. 13523-13535, 2019, doi: 10.1109/ACCESS.2019.2893601.

10. Z. Zhao, J. Wang, H. Sun, Y. Liu, Z. Fan and F. Xuan, "What Factors Influence Online Product Sales? Online Reviews, Review System Curation, Online Promotional Marketing and Seller Guarantees Analysis," in IEEE Access, vol. 8, pp. 3920-3931, 2020, doi: 10.1109/ACCESS.2019.2963047.