ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

SENTIMENT ANALYSIS AND VISUALIZATION OF CUSTOMER REVIEWS

Ms.Apurva Vyankatesh Gundla [*], Prof. Manisha S. Otari

[*] Department of Computer Engineering, N.K.Orchid College of Engineering, Maharashtra, India

ABSTRACT

The Web has become an excellent source for gathering consumer opinions. There are now numerous Web sites containing such opinions, e.g., customer reviews of products, forums, discussion groups, and blogs. All these reviews are the opinions of people all over the world about different products. With the growing availability and popularity of opinion-rich resources such as review forums for the product sold online, choosing the right product from a large number of products have become difficult for the user. These sources are underutilized both by consumers and businesses due to their unstructured nature, serial presentation, limited search tools, and low ratio of useful information to the overall amount of data. The proposed system illustrates visual analysis system that performs sentiment analysis and derives insight from a collection of online reviews of products from customer. Effective visual analysis of online customer opinions is needed, as it has a significant impact on building a successful business and helps the customers in decision making process. This paper presents background Study of Sentiment Analysis or Opinion Mining and gives overview of proposed methodology with insights into past research work.

KEYWORDS: Opinion Mining, Sentiment Analysis, Customer Reviews, Opinions.

INTRODUCTION

Sentiment Analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis aims to determine the attitude of a speaker or a consumer with respect to some topic or product. [1]Sentiment analysis, also called *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. [2]
It involves techniques from different disciplines like information retrieval, Natural Language Processing and Data Mining. Sentiment Analysis is about extracting the opinions or sentiments when given a piece of text. "What other people think" has always been an important piece of information for most of us during the decision-making process. [3]Opinions or sentiments are central to almost all human activities and are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are, to a considerable degree, conditioned upon how others see and evaluate the world. Whenever we need to make a decision, we want to know others' opinions. [2]
An Opinion is a judgment or belief a majority of people formed about a specific thing, not necessarily based on fact/knowledge. Opinion generally refers to what a person thinks about something or opinion is a subjective belief, and the result of emotion or facts interpretation. [4]
In the real world, businesses and organizations always want to find consumer or public opinions about their products and services. Individual consumers also want to know the opinions of existing users of a product before purchasing it, and others' opinions about political candidates before making a voting decision in a political election. In the past, when an individual needed opinions, he/she asked friends and family. When an organization or a business needed public or consumer opinions, it conducted surveys, opinion polls, and focus groups. Acquiring public and consumer opinions has long been a huge business itself for marketing, public relations, and political campaign companies.This unique feature plays a vital role in determining on matters that have financial, medical, social or other implications. Seeking second or third or many more opinions have fuelled the interest of researchers in the field of sentiment mining. [5]
Many reviews are long, which makes it hard for a potential customer to read them to make an informed decision on whether to purchase the product. If he/she only reads a few reviews, he/she only gets a biased view. The large number of reviews also makes it hard for product manufacturers or businesses to keep track of customer opinions and sentiments on their products and services. [6] The substantial gathering of opinions on the Web makes it extremely tough to get helpful data effectively. Perusing all reviews and emotions to settle on an educated choice is a much time taking task. Perusing distinctive and potentially even conflicting opinions composed by diverse

commentators may make organizations, users and customers more confused. It is thus highly desirable to produce a visualization of reviews.The proposed system aims to let users gain useful information for decision making as quickly and as effortlessly as possible, by transforming large collections of reviews text into visualizations that provide the same conceptual understanding that would otherwise require the reading through the whole text collection.

MOTIVATION AND RELATED WORK

Online reviews have became an important source of information for both producers and consumers, with companies trying to better understand customer-provided feedback on products and brands, and individual users looking for information to support their everyday purchasing decisions. Given the widespread use of computers and mobile devices, most of which are connected to the Internet, more and more people are sharing their thoughts, feelings, and experiences. This growing amount of online opinionated information has led to the rapid development of the field of sentiment analysis, which focuses on the identification of opinions, emotions, evaluations, and judgments, along with their polarity positive or negative.

Related Work:

Jeonghee Yi and Nasukawa[7] first used the term sentiment analysis in their paper Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques that *i)*extracts topic-specific features, *ii)* extracts sentiment of each sentiment-bearing phrase, *iii)* makes (topic| feature, sentiment) association.

Minqing Hu and Bing Liu [8] studied the problem of feature-based opinion summarization of customer reviews of products sold online. The task is to identify the features of the product that customers have expressed opinions on (called *opinion features*) and rank the features according to their frequencies that they appear in the reviews. For each feature, identify how many customer reviews have positive or negative opinions. The specific reviews that express these opinions are attached to the feature. This facilitates browsing of the reviews by potential customers.

Xiaowen Ding, Bing Liu and Philip S. Yu [6] represented a holistic lexicon-based approach. Given a set of product features of a product, we want to accurately identify the semantic orientations of opinions expressed on each product feature by each reviewer. Semantic orientation means whether the opinion is positive, negative or neutral.

Bing Liu, Minqing Hu and Junsheng Cheng [9] proposed an analysis system with a visual component to compare consumer opinions of different products. The system is called *Opinion Observer*. With a glance of its visualization, the user can clearly see the strengths and weaknesses of each product in the minds of consumers.

Zhu et al., [10] proposed aspect based opinion polling from free form textual customers reviews. The aspect related terms used for aspect identification was learnt using a multi-aspect bootstrapping method.

Kamal et al [11] implemented a rule based system to mine product features, opinions and their reliability scores. The proposed system uses linguistic and semantic analysis of text to mine the feature opinion pairs from review documents.

Draper and Riesenfeld [12] developed an interactive visualization system to allow users to visually construct queries on large tabular data sets and view results in real time.

Morinaga et al. [13] suggested a 2D scatter plot called positioning map to show the group of positive or negative sentences.

Subjectivity/ Objectivity Classification:

The task of determining whether a sentence is subjective or objective is called subjectivity classification. The text pieces may or may not contain useful opinions or comments. The subjective sentences are the relevant texts, and the objective sentences are the irrelevant texts. So we must sort out the sentences that are useful for us and those which are not. An example subjective sentence is “I like iPhone.”

An objective sentence presents some factual information about the world, while a subjective sentence expresses some personal feelings, views, or beliefs. A subjective sentence may not express any sentiment. Objective sentences can imply opinions or sentiments due to desirable and undesirable facts. An example objective sentence is “iPhone is an Apple product.”

PROPOSED APPROACH AND IMPLEMENTATION

Sentiment analysis is the process of extracting sentiment from fragments of text. As people leave on the Web their opinions or reviews on products and services they have used, it has becomes important to develop methods for sentiment analysis and their corresponding visualization. Such opinions and sentiment analysis have an increasing influence in decision making.Most of the existing methods are processing the reviews in terms of positive and negative comments and rate them in the order of positivity and negativity or summarizes the review information based on features in review sentences.

Sentiment Analysis and Visualization of online reviews of customers is necessary for products. Since a "picture is worth a million words", visualization helps the audience quickly absorb and interpret the presented data. Primary goal of visualization is to communicate information clearly and efficiently to users via the statistical graphics, plots, information graphics, tables, and charts selected. Effective visualization helps users in analyzing and reasoning about data and evidence. It makes complex data more accessible, understandable and usable. As a result, data visualization enables you to present a considerably larger amount of data in comparison to the textual format. The viewer understands what you are trying to say at a first sight. [14] Architecture of proposed system is shown below in figure.

User selects the product and model details from list of products. Reviews of selected product are retrieved online from desired number of e-commerce websites. Stop words of retrieved reviews are removed. Then each word in review sentence tagged with its part- of- speech (such as noun, adjective, adverb, verb etc.). In feature extraction, product features are extracted from each sentence. Product features are generally nouns, so each noun is extracted from sentence. In polarity identification, semantic orientation of each opinion word is identified. Semantic orientation means identifying whether opinion word is expressing positive opinion, negative opinion or neutral opinion. Each extracted opinions contains positive, negative and neutral for each feature. Finally output is visualized in feature wise form of bar charts, line graph.

Proposed Approach:

Proposed approach consist of following modules

Review Extraction Module:

Review Extraction is extracting review information online i.e. Review details of particular product from review sites and e-commerce sites. In proposed system to extract relevant information from websites HTML Agility Pack (HAP), a free, open-source library designed to simplify reading from and writing to HTML documents. It's a process to access external website information (the information must be public – public data) and processing it as required. This is an agile HTML parser that builds a read/write DOM and supports plain XPATH or XSLT (you actually don't HAVE to understand XPATH nor XSLT to use it, don't worry...). It is a .NET code library that allows you to parse "out of the web" HTML files.

POS Tagging Module:

One special application of natural language processing is determining the part of speech of each word in a sentence, known as part-of-speech (POS) tagging. The reason why POS tagging is so important to information extraction is the fact that each category plays a specific role within a sentence. Nouns give names to objects, beings or entities from our world. An adjective qualifies or describes nouns. Also some adverbs can play pretty much the same role as an adjective. In proposed system to tag the extracted review sentence, Apache OpenNLP tagger which is freely available is used. The Apache OpenNLP library is a machine learning based toolkit for processing of natural language text.

Sentiment Analysis Module:

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. A common use case for this technology is to discover how people feel about a particular topic.

The Sentiment Analysis API analyzes text to return the sentiment as positive, negative or neutral. It contains a dictionary of positive and negative words of different types, and defines patterns that describe how to combine these words to form positive and negative phrases. The Sentiment Analysis API takes input for analysis. The API splits the input into entities, which describe different part of the input with a particular sentiment. Each of the sentiments extracted gives information about sentiment, topic and score of the statement.

The output of API is in JSON format and sentiment is about any topic that detected in input. Many times, it contains non-relevant sentiment topics than feature list of product we have. We process output of an API in our developed system. First our code reads the output in required format from different reviews. Then filter out the output features listed as per our requirement. System calculates the average score for different features and decides polarity accordingly. After, calculating the average score, deciding polarity, system saves the reviews for presentation to user in terms of feature – review – polarity. While, score calculated was used for visual representation of user reviews.

Visulaization Module:

The main purpose of this module is to visualize the analyzed text results in graphical presentation. Data visualization is the presentation of data in a pictorial or graphical format. For centuries, people have depended on visual representations such as charts and maps to understand information more easily and quickly. The human perceptual system is highly attuned to images, and visual representations can communicate some kinds of information more rapidly and effectively than text. Visualizations are the single easiest way for our brains to receive and interpret

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

large amounts of information. The purpose of data visualization is to simplify data values, promote the understanding of them, and communicate important concepts and ideas.

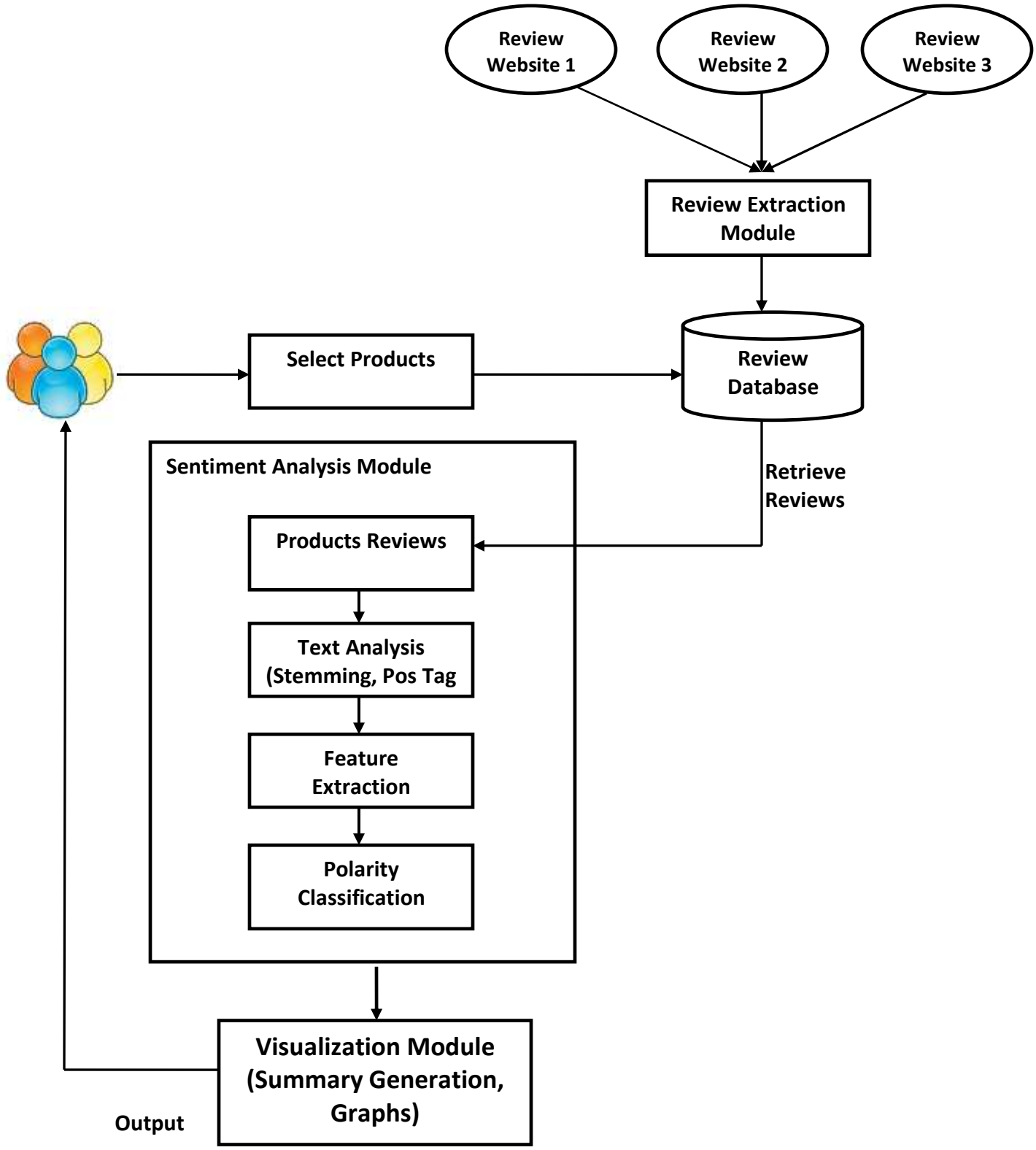


Fig: 1. Architecture of Proposed System.

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

ALGORITHMS

Algorithm of Review Extraction

Input: Links of websites from which reviews is to be extracted.
Output: : Reviews of particular product.
Steps:

1. HTMLAgilityPack is used to fetch/extract web page contents. HTMLAgilityPack is a tag based extraction system which reads the web page contents for given tag. Such as, the contents starting with the tag.
2. GetWebSiteContents(string url) method to fetch web contents from the webpage containing reviews for selected product.
3. As, reviews are displayed using paragraph, such as 'p' tag, use 'p' tag to read data.
4. The tag contains inner html and inner text data formats while it has attributes like color, text, font, etc.
5. Inner html format contains data along with style properties and any child tags while inner text contains only plain text data.
6. As we need plain text review, we use inner text data format to fetch contents from web page.
7. For all websites links under consideration, 'p' tag nodes are collected and read.
8. Now first/next 'p' tag from collection will read.
9. If tag is not null, then system will read inner text such that review for current 'p' tag and append to string builder using separator '#' from previous review. Then it goes for next tag in tag collection.
10. The steps 9 and 10 are repeated until all tags from tag collection are read out.

Algorithm for POS Tagging

Input: Extracted review sentences.
Output: : Parts of speech for each word in review sentence.
Steps:

1. For POS tagging, first identify the task, train and build the model.
2. For training, test.pos and train.pos data files are used.
3. Remove all special characters occurrences and consider one single statement to tag.
4. SeparatorList is a list which contains characters to be used for statement separation.
5. Initiate the definitions/descriptions of tags to be bind and used in output
6. Training of an application is performed. The output of training is stored in hash map (Hash map is like a temp database which stores large amount of data which is useful throughout application) after training, method for tagging real time data was called.
7. SplitSentences method is used to detect statements from paragraph under consideration using training file EnglishSD.nbin.
8. TokenizeSentence method is called to tokenize and tags the words in sentence. The method initiates the variable with tokenize information from EnglishTok.nbin file.
9. After tokenizing information such as tagging of words, the output will get attached to end of tree structure.
10. Dynamic array type 'List' is used to add output, the new node of type list is initiated which is attached to end of tree in form of node branch.
11. To present the POS information of statements in output, each node of tree is get read by module.
12. If node contains word and tag information then look for tag description in dictionary. If description is available, add it to tag information of word
13. Merge the token and tag information in a single variable to be return to calling method

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

Algorithm: Sentiment Analysis
Input: Extracted Reviews.
Output: Polarity negative, positive, neutral..
Steps:

1. System goes through training set of POS Tagger to judge the words in statements.
2. The word types define the weight of word in terms of sentiment. Consider a single statement for analysis
3. POS tagger module is used to classify the statements into subjective statement and objective statement.
4. The system contains a dictionary of positive and negative words of different types, and defines patterns that describe how to combine these words to form positive and negative phrases.
5. System analyzes text to return the sentiment as positive, negative or neutral and decides the polarity of statement
6. For each of the sentiments extracted gives information about sentiment, topic and score of the statement
7. The output of sentiment analysis is a sentiment about any topic that detected in input.
8. System process output to filter out the output features listed as per our requirement.
9. System calculates the average score for different features and decides polarity accordingly.
10. Saves the reviews for presentation to user in terms of feature – review – polarity. While, score calculated was used for visual representation of user reviews.

Algorithm: Visualization Module
Input: Evaluated scrores of product features
Output: Bar graph, spline chart, speedometer.
Steps:

1. The score for features of selected product are displayed using spline chart and bar chart.
2. Object Data Source is used to bind chart data such that axis(x and y) data and corresponding data from database.
3. "SelectMethod" of Object Data Source is used to call data generation method from class file ChartData.cs.
4. "GetData()" of return data type "List" generates chart data. Declare "List" data type variable "data"
5. The data type "List" is a dynamic array data type whose array size get increases as we add data to end of array.
6. Count total available records for features of selected product-model.
7. Initiate an array of features to selected features, count and calculate average polarity/score for each feature from feature array.
8. Convert average polarity into corresponding percentage value. Append feature and corresponding average polarity/score to data variable "data"
9. Read all values, return variable "data" to calling method, then Object Data Source will displayed graph.

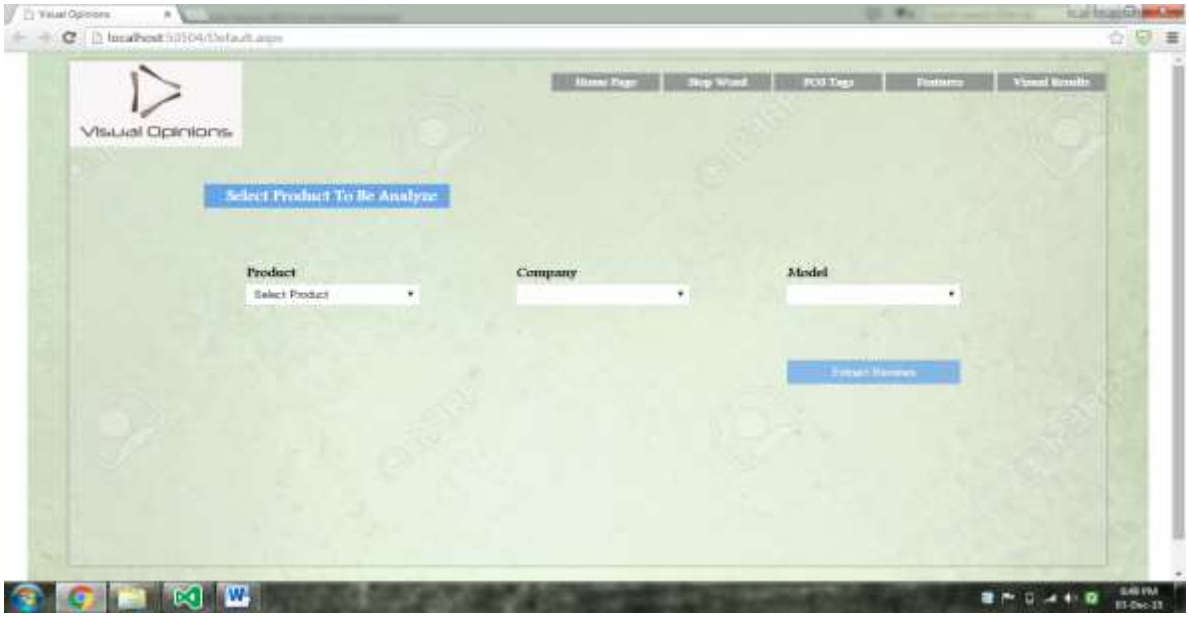
ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

EXPERIMENTAL SETUP

In the experimental phase of the system consist of four tabs namely Stop Word, POS tags, Features and Visual results. User first selects the product to review and analyze. User selects the details of product like company and model name of the product to be reviewed and clicks on extract reviews button. After selecting the product details by user system extracts online reviews from one of the ecommerce websites. The links of the websites are already saved in database. Extracted reviews are displayed to user in table form. The stop word tab displays the reviews with all the stop words being eliminated or removed. The POS Tag displays the parts of speech tags such as noun, adjective, verb for each word in the review sentence. The features tab shows the extracted features from the reviews along with the opinions on the specific feature and classified polarity for each feature and estimated score. The visual results tab presents the feature wise graphical view of the reviews in the form of spline chart, speedometer and bar graph.

The result of the system consists of following screenshots.

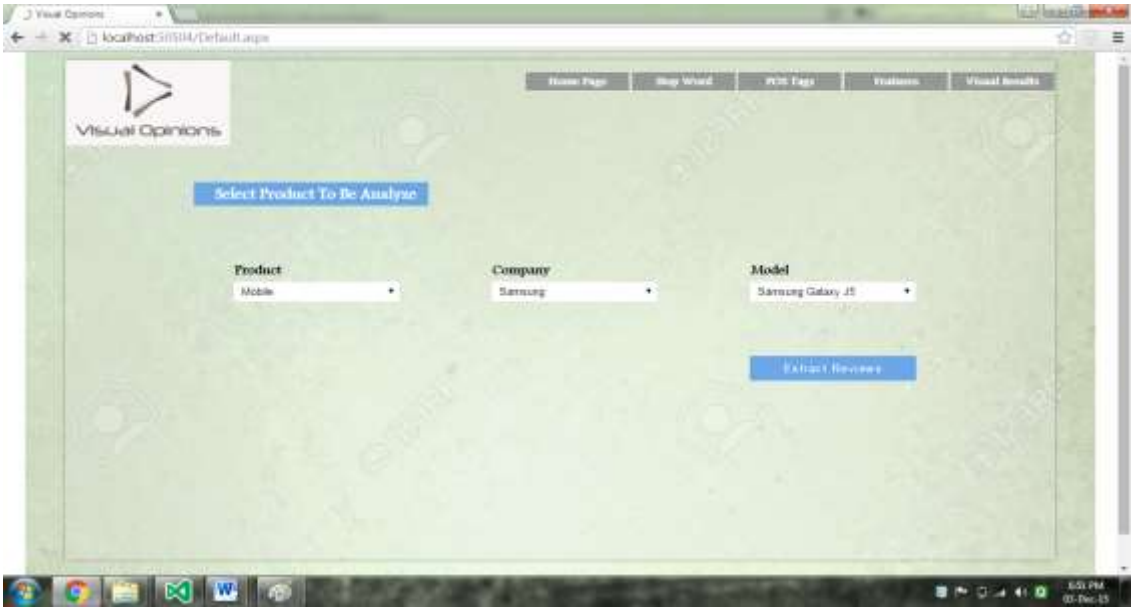
Screenshot 1 gives the overall view of the system. Results of the system are displayed in four tabs namely Stop words, POS tag, Features and Visual results. It consist three dropdown lists to select the product, company and model name.



Screenshot 1: Design of proposed system

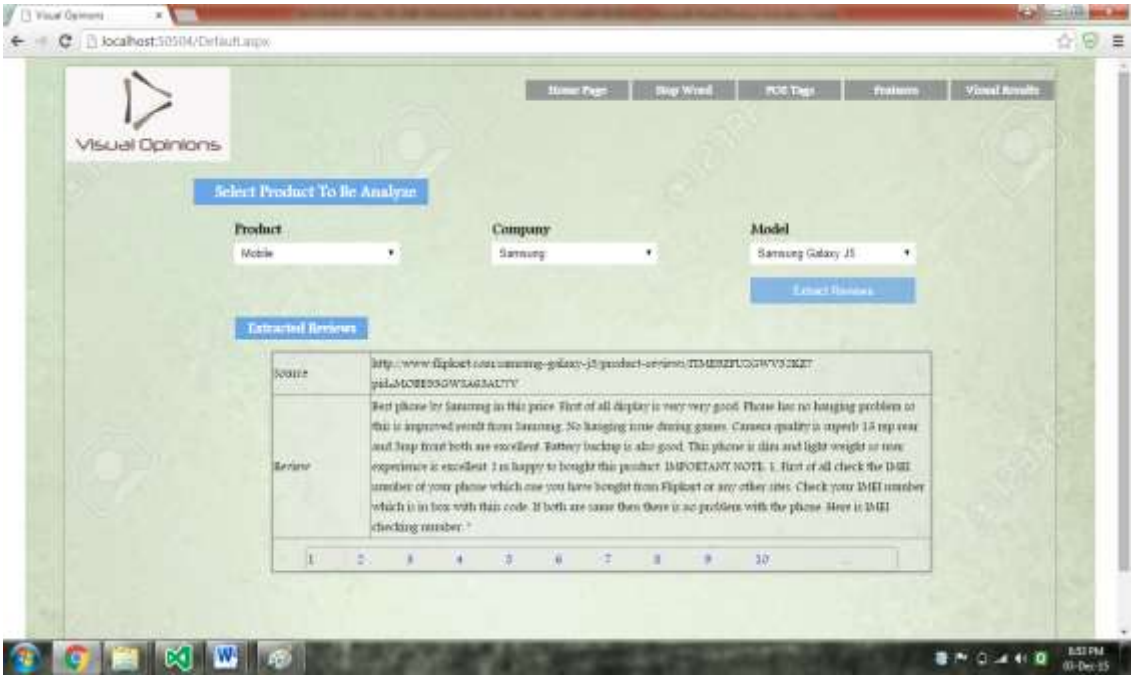
In Screenshot 2, the user selects the product and its details like company and model name to analyze the reviews and clicks on extract reviews button.

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785



Screenshot 2: User selects the product details to review.

Screenshot 3 shows the result of online extracted reviews from e-commerce websites whose links are saved in the database.



Screenshot 3: Online Extracted Reviews.

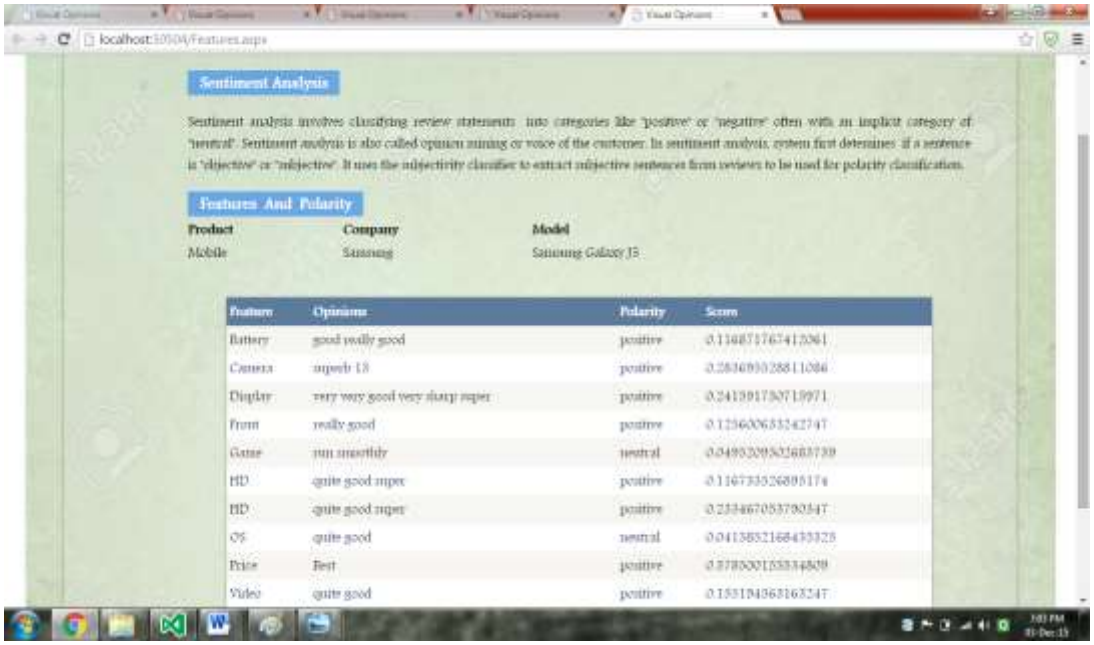
Screenshot 4 shows the result of reviews of which all stop words are eliminated, user has to click on stop word tab to view the result which is to the right side of the system.

Screenshot 4: Stop words eliminated from reviews

Screenshot 5 shows the result of POS tags tab, parts of speech tag is assigned to each word in the review sentence.

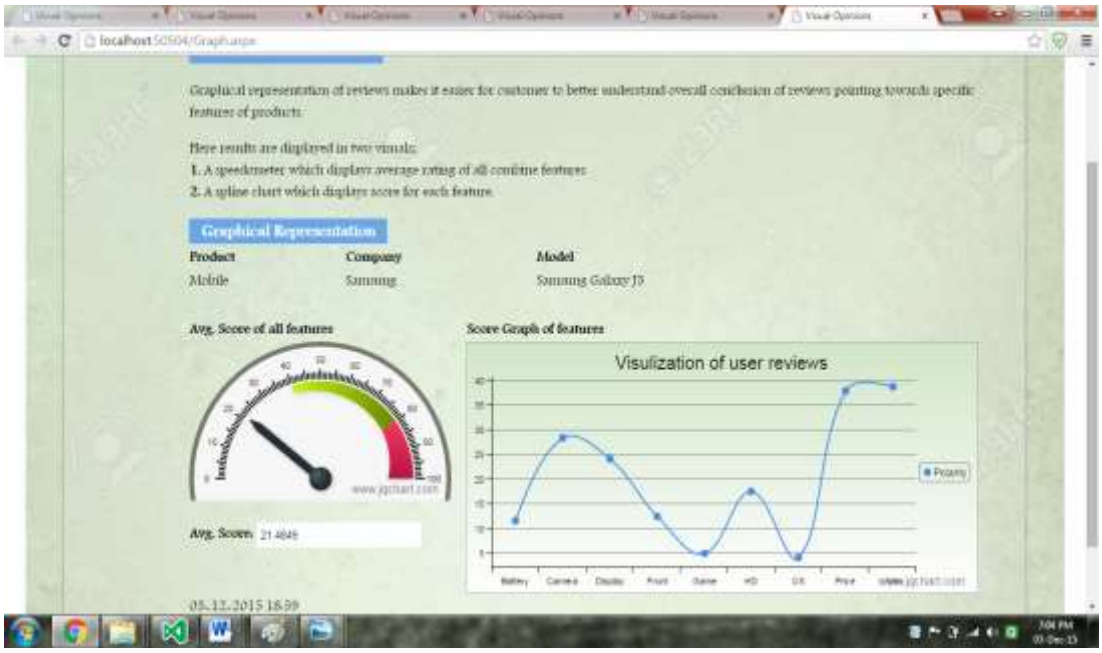
Screenshot 5: POS tags of each word in the review sentence.

Screenshot 6 displays extracted features from the review sentences, customer's opinions on the features, classified polarity of each feature and evaluated score of features.



Screenshot 6: Features, opinions on the features, polarity and scores of features.

Screenshot 7 gives feature wise visual representation of the product reviews in the form of graphs like spline chart, speedometer etc.



Screenshot 7: Visual representation of reviews in the form of graphs.

CONCLUSION

Sentiment Analysis has become a fascinating research area due to the availability of a huge volume of user-generated content in review sites, forums and blogs Applying Sentiment analysis to mine the huge amount of unstructured data has become an important research problem. Proposed System presents a web based sentiment

analysis and visualization tool for online customer reviews for products purchased on e-commerce sites. The system is capable of extracting the online reviews from e-commerce websites. The system tags the individual review sentence using parts of speech tagger which is necessary to identify features of products and opinion words. System analyzes the sentences to decide the polarity and score which can be positive, negative and neutral for each feature. The system converts the unstructured text into feature-wise visual representation like bar graphs, spline chart etc., based on calculated polarity scores. The proposed system helps users gain useful information for decision making as quickly and as effortlessly as possible, and provides the conceptual understanding that would otherwise require the reading through the whole text collection.

REFERENCES

[1] https://en.wikipedia.org/wiki/Sentiment_analysis
[2] Sentiment Analysis and Opinion Mining April 22, 2012 Bing Liu.
[3] Opinion mining and sentiment analysis Bo Pang1 and Lillian Lee2.
[4] Padmaja, S., & Fatima, S. S. (2013). Opinion Mining and Sentiment Analysis–An Assessment of Peoples' Belief: A Survey. *International Journal*.
[5] A Survey of Classification Methods and Applications for Sentiment Analysis 1M.Govindarajan , 2,Romina M.
[6] Philip S. Yu Ding Xiaowen, Liu Bing. A holistic lexicon- based approach to opinion mining. WSDM'08, 2008.
[7] Jeonghee Yi, Nasukawa, Bunescu , Niblack, W., "Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques" T hird IEEE International Conference on , 10.1109/ICDM.2003.1250949, 19-22 Nov. 2003.
[8] M. Hu and B. Liu. Mining opinion features in customer reviews. In AAAI'04: Proceedings of the 19th national conference on Artificial intelligence pages 755–760.
[9] B. Liu,M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In International Conference on World Wide Web, 2005.
[10] Zhu, Jingbo Wang, Huizhen Zhu, Muhua Tsou, Benjamin K. Ma, Matthew, "Aspect-Based Opinion Polling from Customer Reviews", IEEE Transactions on Affective Computing, Volume: 2,Issue:1 On page(s): 37. Jan-June 2011.
[11] A Kamal, M. Abulaish and T. Anwar, "Mining feature -opinion pairs and their reliability scores from web opinion sources," WIMS '12, June 13-15, 2012 Craiova, Romania.
[12] G. Draper and R. Riesenfeld. Who votes for what? a visual query language for opinion data. IEEE Transactions on Visualization and Computer Graphics, 14(6):1197–1204, 2008.
[13] S.Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. Mining product reputations on the web. In ACM SIGKDD international conference on Knowledge discovery and data mining, pages 341–349, 2002.
[14] http://www.uauug.org.uk/what-are-the-advantages-of-data-visualisation.html

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

AUTHOR BIBLIOGRAPHY

	**Miss Apurva Vyankatesh Gundla** She has received B.E Degree in Information Technology from University of Solapur, Maharashtra, India and pursuing the M.E. degree in Computer Science and Engineering in Nagesh Karajagi Orchid College of Engg. & Technology, Solapur, India. She is doing her dissertation work under the guidance Prof. M. S. Otari , Assistant Professor at Nagesh Karajagi Orchid College of Engg. & Technology, Solapur, Maharashtra, India..
	**Prof. M. S. Otari** She is an Assistant Professor in Nagesh Karajagi Orchid College of Engineering and Technology, Solapur, Maharashtra, India. Her research area includes Machine learning, Natural Laguage Processing, Web Technology, networking.