A Novel to Project Customer Feelings via

M. Maragatharajan, N. S. Aarthy, K. Mahalakshmi, M. Muthukumar, C. Balasubramanian

Abstract—When buying a product people searches for several options on the internet so that they acquire only the best ones from the market. To assure this, people go to the internet and search for reviews. Based on the reviews the product will be chosen by the customers. When we think about this, there will be a lot of confusion which is the best one. To prevent this, reviews are rated and then they are categorized. Based on the classification the product is named as a good or bad product. To address the problem of classification we use SVM classification. SVM algorithm is chosen since it is easy and user-friendly. The Main aim of this work is finding the opinion of the product on marketing based on their reviews. Online reviews can help them improve the quality of services and products. However, it is an ambitious issue to take on valuable online reviews by realizing the review sentiment. This paper will dynamically determine the customer opinion about the specific product whether the customer will have an optimistic or bad opinion about the product. Online evaluations have transformed the vital resource of data for users prior to making an up-to-date purchase decision. Identifying evaluations is cooperative to check and run early endorsement and also reviewers are likely to be the buyers of a product.

Keywords— Aspect-based, opinions, reviews, feelings, analysis, and outcome

I. INTRODUCTION

As there is a quick growth of commercialism, large integer of clients can at present contribute their thoughts around numerous kinds of products in conversations, commercial websites, their blog, or a review article. As a effect, the abstraction of product reviews available on the internet is increasing rapidly. Already, the number of available reviews makes it unreasonable to future clients who construe them completely also distinguish the promise belief just about the trade good. Opinions are those words which come out either in verbal or written form from a customer who has already used the product.

These opinions may vary from individual to individual so it becomes difficult for a first time user to identify the pros and cons of manufactured goods Therefore, instinctive opinion recognition and summering techniques have risen to assist people to make a learned decision with the Machine learning algorithms. Machine learning is a logical train that investigates the development and investigation of controls that can gain from data.

Revised Manuscript Received on December 16, 2019.

M. Maragatharajan, Assistant Professor, Department of Information Technology, Kalasalingam Academy of research and education, Krishnankoil, India.

N. S. Aarthy, Department of Information Technology, Kalasalingam Academy of research and education, Krishnankoil, India.

K. Mahalakshmi, Department of Information Technology, Kalasalingam Academy of research and education, Krishnankoil, India.

M. Muthukumar, Department of Information Technology, Kalasalingam Academy of research and education, Krishnankoil, India.

C. Balasubramanian, Assistant Professor, Department of Computer science and engineering, Kalasalingam Academy of research and education, Krishnankoil, India

Exactly, the machine learning algorithms can be segregated into the following ways: Supervised, Unsupervised, Semi-Supervised. The methods involved in our work are data processing, tokenization, stop words removal, Term Frequency Construction, Classification of Reviews and prediction.

In this paper, we have planned a system that will help us to segregate the reviews of the people and help others to take the right decisions when buying a product. This novel approach is designed to make the work of people even simpler. The first step is to select the data set and then we load it. The second step is data processing. In this process, we are noticing, modifying or eliminating corrupt or mistaken records from the data set. The records which provide the incorrect clustering results are detected and removed from the dataset. Also in this process, we are going to delete the records that contain empty fields. The tool that we are planning to use is the outlier technique. The third step is tokenization. Tokenization the act of breaking up a classification of filaments into fragments such as words, keywords, phrases, symbols and other elements called tokens. Individual words, phrases or even whole sentences can also be labeled as tokens. The fourth step is to stemming. Stemming is the process of converting the words of a sentence to its non-changing portions. The Porter stemming algorithm is used for stemming the words. The fifth step is the term frequency constructor. Term frequency is used in joining with data retrieval and shows how frequently a term occurs in a document. The sixth step is clustering. The seventh step is the classification. For classification the algorithm chosen is SVM. The last and the final step is the projection. In this step, we project the value of the product.

II. THE EXISTING SYSTEM

In the existing system, the reviews of the product were only the definite quantity of accessible reviews making them unrealistic when it comes for our potential clients to construe those reviews also separate the compromise thoughts close to the goods [1-3]. Consequently, public opinion perception, summarizing techniques hold up for assisting groups to sort a choice. Public opinion detection can as well be called as sentiment analysis. Every time somebody tries to see what other people think about anything on the web, the reply is a huge amount of data, which makes it difficult to find useful information easily. For organizations, tracking feedback can help to calculate the height of fulfillment and make best manufacturing and selling decisions. Therefore, systems that can mechanically sum-up documents are sentiment analysis abstract treasured biased data through the unprocessed textual matter of the review.

Public opinion detection is classified into three: record-level, unverifiable/objective recognition, regarding views opinion abstraction. Opinion classification is the most largely searched subject, which divides a opinions into positive and negative.Unverifiable/objective determination determines objective string of words which consider opinions [4]. Nevertheless, the assorts is excessively harsh for almost all of the live utilities since settlement cannot be done on accurately what each person has interests in. Abstracting data in consumer opinions sorted on opinions is the most impressive manner for assisting clients resourcefully digests the vast quantity of on-hand data. In this process, there are no algorithms used to classify the data and get the output. The classification is done in a random manner and the result obtained is in a coarse manner. A supporting copy on a particular article does not signify that the writer like or dislike every single aspect of the article [5]. It is not suitable to categorize and conclude the emotion on individual products.

III. THE PROPOSED SYSTEM

The proposed system is used to project the exact outcome of the product based on the reviews of the people. This involves many processes as mentioned above. The first step is data selection and loading. For this paper, we have chosen to use an Amazon dataset which has about 1000 records. Then we load them. After loading them we start the process of data processing. In the process of data processing we remove the null data, incomplete data, and incorrect form of data [6]. We try to convert them into the form that can be loaded into the machine. This process is done by outlier technique. Outlier technique is an observation that is distant from observations. We use this technique a couple of times to make sure we get nearly the same result each and every time. This is also done to ensure that the error rate is less. After loading them into the machine we ought to do the process of tokenization. This is the process of grouping similar sort of words into a group and giving them tokens.

The tokens are labels that will help us to identify the words to which they belong to. By identifying the group to which they belong to it is used by the users for easy classification in the later stages of the paper. The next step is that we are going to remove the stop words. This process is called stop words removal. This involves the process of removing unwanted articles, prepositions, etc. This helps in bringing the reviews to that point rather than being vague [7-9]. The next process is stemming. The process of stemming is to convert the words of a sentence to its non-changing portions. This means that when all the unwanted articles are removed we only get the gist of the review. We have proposed to use the Porter stemmer algorithm. We have chosen to use this since it gives better results, much easier to use and understand. The porter stemmer algorithm converts the words into its root words. The next step of progress it to do a process called term frequency construction. Here we show how frequently a word is being used. The most important work begins here. This process is called classification [10].In this process, we classify the reviews. To do this we use the SVM classification. We use SVM because it is easy to use and understand the process that is going on with the data. In the last step, we do the projection which is used by the customers to buy their desired product. Fig 1 illustrates the architecture of the proposed model:

A. DATA PROCESSING

Data processing is a vital move in the data mining process. If a large amount unrelated, outmoded data are there and unpredictable data in the data set, then our process becomes more difficult. Data collecting methods are often loosely controlled data, unfilled data, missing values, etc. Data that has not been carefully filtered for such problems can produce misleading results.

B. TOKENIZATION

Tokenizationis the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can also be individual words, phrases or even whole sentences. The process of tokenization is done to make sure that tokens are used to count the number of times a token has been repeated [11]. Based on the number of times a token has been repeated the importance of the word is known to the user. Some of the common known tokens can be of identifiers, separators, operators, literals, etc. By doing this we can make sure that accurate results can be applied

C. STOP WORDS REMOVAL

Stop words removal means removing unwanted words. Stop words are natural language words which include a very little meaning such as "and", "the", "a", "an", and similar words. The words are detected and it's removed. This is done to ensure that the reviews are precise and to the point.

D. STEMMING

Stemming is the procedure of shrinking big terms to their word stem. The stem word need not be the same as the morphological root of the word. It is usually sufficient that related words record to the same stem. A computer program that stems word may be called a stemming program, stemming algorithm, or stemmer. There are many types of stemming. One of which is the Porter stemmer algorithm. Suffix removal does not rely on a look-up table that consists of root form relations. Most common examples are: when a word has 'ed' in its end then the 'ed' is removed. Similarly the same is done for 'ly' and 'ing'. In porter stemmer algorithm also the same is done. In porter stemmer algorithm the common errors are also noted and converted into a useful word [12-14]. These errors may include spelling mistakes, grammatical errors, and misplaced letters in a word, short-forms of words etc. Here, the root words are store in a separate folder and this folder is referred during the process. This process plays a vital role since in this paper we consider only the root words for reviewing.

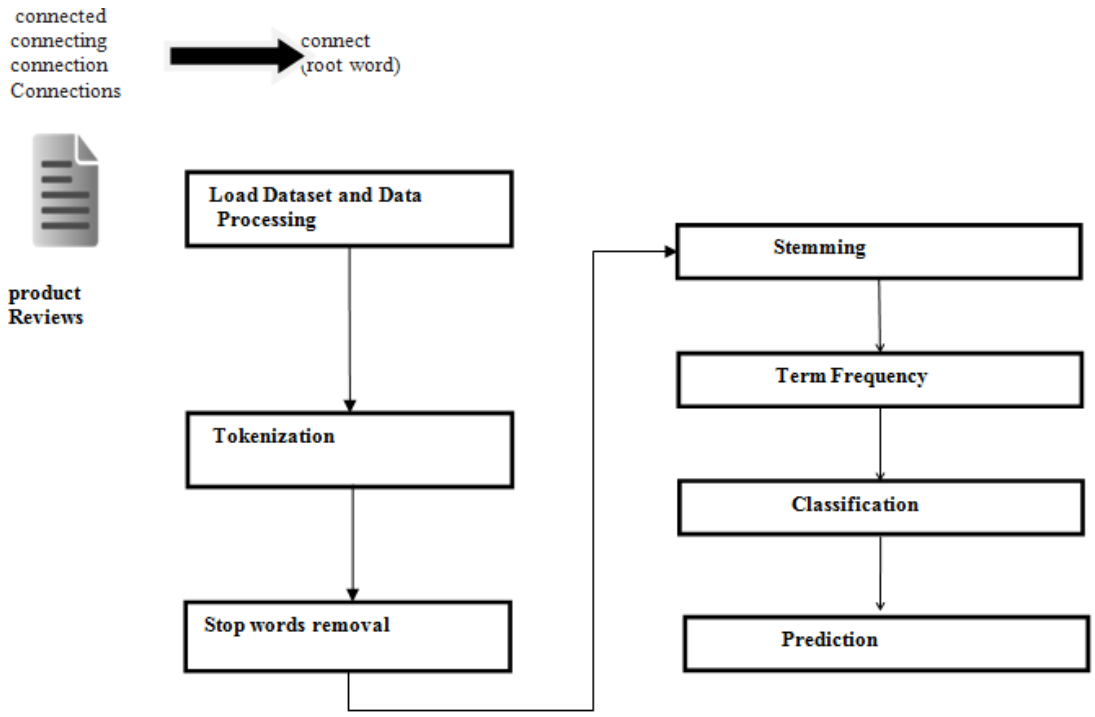

Fig 1 Architecture of Proposed Model

E. Term Frequency Construction

Term frequency is to be used in association with information recovery and shows how frequently an expression occurs in a document. It specifies the importance of a particular word inside the document. This is a quantitative compute which will be used for achieving magnitude of a statement in a record according to how rarely it happens to be views in the record.Opinion of this measure is: If a word appears frequently in a document, then it should be important and we should give that word a high score. But if a word appears in too many other documents, it's probably not a unique identifier; therefore we should assign it as poor rating for the statement.

C. Classification

Classification is used to classify each item in a set of data into one of the predefined set of classes or groups. Classification is a data mining function that assigns bits and pieces in a set to aim categories or classes. The objective of classification is to exactly predict the aim class for each case in the data. In this we are planning to use Support vector machine. Support Vector Machine (SVM) is a supervised machine learning algorithm which can be applied for classification [15-17]. In this algorithm, we design each data item as a point in n-dimensional space where n is the number of attributes we have with the value of each feature being the value of a picky coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes. Support Vectors are simply the coordinates of individual observation. Support Vector Machine is a border which best separates the two classes hyper-plane/ line.

The steps involved are as follows: recognize the hyper-plane, classify two classes, Find the hyper-plane to separate to classes and suggesting a new plane after many tryout and errors. In addition to performing linear classification, SVMs can powerfully do a non-linear classification using what is called the kernel trick. When data is unlabeled, supervised learning is not possible, and an unsupervised learning approach is required, which trials to find the natural grouping of the data to groups, and then map new data to these formed groups.

IV. RESULTS AND DISCUSSION

Based on the data set acquired we have got a certain result. This includes the data set obtained from Amazon for tablets, I-pads, I-pods, speakers and headsets.

Table1 : Dataset received

Item	No. of reviews	No. of sentence	Training set	Test set
Tablet	1000	1200	600	400
ipad	256	500	130	126
ipod	54	100	30	24
Speakers	500	700	400	100
Headsets	750	950	400	350
Charger	200	350	120	80

Based on the above table one can see that even when there are many reviews a product is bought based on personal opinions. This table is also proof that when there are a larger number of opinions we have to train a certain number of data to get results which are near to accurate

Fig 2: Customer Opinion

Based on the above graph which is taken from the data set1 we can come to the conclusion that most of the products are good and only some are bad. We have used a pictorial representation because pictures speak a lot than words. Predictions are made based on the bar graph obtained in the last module. Since this is an application there are few differences from the internet available data processing tools [18-20]. This application's modules can also be designed using other programming languages like c#, python , etc. We have chosen net beans since it is easy and efficient to use. Also in this work, we have used SVM algorithm with the support vectors of classes c1 and c2 for efficient and accurate classification of the data set. In future, a user can also use other algorithms like naive bayers, logistic regression, K-nearest forest, etc for classification and projection of data opinions. This is a combinational work involving programming and algorithms for obtaining a successful output.

V. CONCLUSION:

Our report address customer reviews on products abstracted based on facts of manufactured goods opinions. Also to bring in new methodology which follows a few steps. The data-set is downloaded and the data is processed. After data processing, the obtained dataset is loaded for tokenization, stemming and result generation. Here we have used SVM algorithm since it is easy to handle and easy to modify according to the needs of the user. For future users, one can refer this and choose any other algorithm like naive bayers to obtain the similar output. One can also use python programming to design the same application. Product opinion mining was to be performed and accomplished to obtain relevant results, especially for the first three domains. The scheme possibly intent a new way to goods feature removal in bigger data sets also possibly be useful for difficult work like silent judgment inference, sentiment action possibility.

REFERENCES

[1] G. Salton, A. Singhal, C. Buckley, and M. Mitra, ``Automatic text decomposition Using text segments and text themes," in Proceedings HYPERTEXT, Bethesda, MD, USA, 1996, pp. 53_65.
[2] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell, ``Summarizing text documents: Sentence selection and evaluation metrics," in ProceedingsSIGIR, Berkeley, CA, USA, 1999, pp. 121_128.
[3] C. D. Paice, ``Constructing literature abstracts by computer: Techniques and prospects," Inf. Process. Manage., vol. 26, no. 1, pp. 171_186, 1990.
[4] J. Kupiec, J. O. Pedersen, and F. Chen, ``A trainable document summarizer, "in Proceedings SIGIR, Seattle, WA, USA, 1995, pp. 68_73.
[5] O. Gross, A. Doucet, and H. Toivonen, ``Document summarization based on word associations," in Proceedings SIGIR, Gold Coast, QLD, Australia, 2014,pp. 1023_1026.
[6] K. Dave, S. Lawrence, and D. M. Pennock, ``Mining the peanut gallery: Opinion extraction and semantic classi_cation of product reviews," in Proceedings. WWW, Budapest, Hungary, 2003, pp. 519_528.
[7] C. Whitelaw, N. Garg, and S. Argamon, ``Using appraisal groups for sentiment analysis," in Proc. CIKM, Bremen, Germany, 2005,pp. 625_631.
[8] S. R. Das and M. Y. Chen, ``Yahoo! For Amazon: Sentiment extraction from small talk on the Web," Manage. Sci., vol. 53, no. 9, pp. 1375_1388, 2007.
[9] A. Devitt and K. Ahmad, ``Sentiment polarity identification in financial news: A cohesion-based approach," in Proceedings ACL, Prague, Czech Republic, 2007, pp. 984_991.
[10] B. Liu, ``Sentiment analysis and opinion mining," Synth. Lectures Human Lang. Technol., vol. 5, no. 1, pp. 1_167, 2012.
[11] V. Hatzivassiloglou and J. M. Wiebe, ``Effects of adjective orientation and gradability on sentence subjectivity," in Proceedings COLING, Saarbrücken, Germany, 2000, pp. 299_305.
[12] J. M. Wiebe, ``Learning subjective adjectives from corpora," in Proceedings AAAI/IAAI, Menlo Park, CA, USA, 2000, pp. 735_740.
[13] J.Wiebe, T.Wilson, and M. Bell, ``Identifying collocations for recognizing opinions," in Proc. ACL/EACL, Toulouse, France, 2001, pp. 24_31.
[14] P. D. Turney, ``Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in Proceedings ACL, Philadelphia, PA, USA, 2002, pp. 417_424.
[15] P. D. Turney and M. L. Littman. (2002). ``Unsupervised learning of semantic orientation from a hundred-billion-word corpus." [Online]. Available:https://arxiv.org/abs/cs/0212012
[16] B. Pang, L. Lee, and S. Vaithyanathan, ``Thumbs up? Sentiment classification using machine learning techniques," in Proc. EMNLP, Philadelphia,PA, USA, 2002, pp. 79_86.
[17] A. Esuli and F. Sebastiani, ``SENTIWORDNET: A publicly available lexical resource for opinion mining," in Proc. LREC, Genoa, Italy, 2006,pp. 417_422.
[18] B. Ohana and B. Tierney, ``Sentiment classification of reviews using SentiWordNet," in Proc. IT&T Conf., Dublin, Ireland, 2009.
[19] S. Baccianella, A. Esuli, and F. Sebastiani, ``SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in Proceedings LREC, vol. 10. 2010, pp. 2200_2204.
[20] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, ``Sentiment strength detection in short informal text," J. Assoc. Inf. Sci. Technol.,vol. 61, no. 12, pp. 2544_2558, 2010.

AUTHORS PROFILE

Maragatharajan M received his Bachelor degree in Electronics & Communication Engineering from Anna University by 2007. He has received his Master degree in Information Technology from Kalasalingam University, 2010 and he has completed his ph.D in the area of MANET. He has worked as a Project Associate in TIFAC CORE in Network Engineering, Kalasalingam University from 2007 to 2008. Currently, He is working as an Assistant Professor in the Department of Information Technology, Kalasalingam University. His areas of interest are Ad-hoc Networks, Wireless Networks, and Network Security.

Aarthy NS has completed her Under Graduation in Information Technology from Kalasalingam Academy of Research and Education, Krishnankoil. Being a topper of the department, placed MNCs. Her area of interest is Web Designing and Networking.

Mahalakshmi K completed her Under Graduation in Information Technology from Kalasalingam Academy of Research and Education, Krishnankoil. Her area of interest is Database Management systems and Data Mining.

Muthukumar M completed his Under Graduation in Information Technology from Kalasalingam Academy of Research and Education, Krishnankoil. He has actively engaged in many workshops and seminars. His area of interest includes Networking and Data mining.