Effective and Dynamic Multi Domain Keyword Extraction through Word Vectors

[1]Neelkanth Poosa, [2]Pranav Sai Marla, [3]M. Venu Gopalachari*

[1]*UG Student, Department of IT, Chaitanya Bharathi Institute of Technology, Hyderabad, Telangana, India*

[2]*UG Student, Department of IT, Chaitanya Bharathi Institute of Technology, Hyderabad, Telangana, India**

[3]*Associate Professor, Department of IT, Chaitanya Bharathi Institute of Technology, Hyderabad, Telangana, India*

[1]neelkanthpoosa@gmail.com, [2]pranav.marla@gmail.com, [3]mvenugopalachari_it@cbit.ac.in*

Abstract

Online product reviews are a customer's best resource to judge a product. Accurate and useful reviews for a product can be challenging to find. Reviews can be tainted, biased, inconsistent or not to the point. This leaves room for consumer-centric review analysis techniques. Multi-Domain Keyword Extraction using Word Vectors is a technique that is aimed at providing the customer with reviews from multiple websites, along with detailed analyses on the reviews, to streamline the customer's experience. The reviews are dynamically scraped from multiple e-commerce websites using the product's unique model number. Machine learning is used to accurately identify aspects and key phrases in the reviews, and context-based sentiment analysis is used to establish the average opinion for each keyword. The machine learning algorithms will process the data in the form of word embeddings to accurately find the keywords in large texts. Review trustability phase is a unique approach formulated to identify trustable reviews considering various aspects that defines a review to be credible. Therefore, this method offers e-commerce customers a consolidated review analysis technique.

Keywords: *Aspect Based Sentiment Analysis(ABSA), Web Scraping, Model Number, Review Trustability, Cold start, Dependency Parse Tree, Aspects.*

1. Introduction

Finding informative reviews for a product online can be a time staking affair. A user needs to browse through a number of reviews over many websites to get an idea of the strengths and flaws of a product. As customers come across tons of reviews on social media, a more accurate representation of all reviews combined is the real deal what they wish to know. Aiding the user with trustable reviews based on certain aspects would also benefit in the decision-making. Finding a way to efficiently summarize review data would not only help customers make smarter decisions, it would make the market as a whole more sensitive to quality. Having an efficient review analyzer in its toolbox would also enable organizations to get quick feedback which can be used to improve their services. Hence there's a space in the market for an efficient keyword and polarity extraction technique that would aid both the customers and organizations. The applications of the keyword extraction and polarity of sentiments are numerous and across many fields. The primary applications involve product review analysis, market analysis and customer opinion analysis. The fields that they can be used for range from electronics to services to e-commerce.

Though there were attempts made by e-commerce sites like Amazon and Flipkart to represent the features of a product, they are often lesser in number and don't hold the credibility of major aspects since they analyze the reviews of their corresponding domain only. This leads to a problem called as cold start, which affects the quality of information retrieval, ignited by the lesser data available in terms of size or users.

This paper addresses the above problems with a system that efficiently gathers reviews of a product from all the three popular e-commerce domains like Amazon, Flipkart and Snapdeal, and extracts the features of the product discussed along with their corresponding adjectives, keeping various other structural text discrepancies checked. The sentiment polarities of the aspects extracted are calculated by the corresponding adjectives that are extracted. In addition to that, the system analyses a review and allots a trustability score based on several viable considerations that make a review trustable by the customers and those which provide significant insights on the product. The process initially begins with the search of product title which initiates the process of scraping from multiple websites that dynamically extracts the reviews and other metadata, and further feeds it to another process that analyses the data and extracts the aspect-adjective pairs with the help of extraction rules formulated. Further, as mentioned earlier, the sentiment polarities are calculated after undergoing word vector and clustering methodologies to derive final features.

The remaining contents of the paper are organized as follows: Section II deals with related work for the proposed work, Section III deals with the system architecture and the methodologies, Section IV deals with results and discussions of the proposed system and Section V emphasizes on the future scope of the system and concludes the research.

2. Related Work

The entire process can be split into the multi-domain scraping part and the aspect extraction part. As for the scraping of reviews from websites, previous approaches have depending on a sole website to extract them from. Multi-Domain scraping runs into the problem of identification of a product, and accessing reviews which are complicated to access using simple HTML parsers. These problems are overcome by using automated browsing tools like Selenium which can handle browsing tasks and by searching for products based on unique identification codes for each product by means of its model number.

As for the aspect extraction part, previous approaches have used a multitude of techniques to facilitate it including parsing, named entity recognizer [1], bag-of-words [2], semantic analysis [3] as well as domain-dependent ones, like word clusters. There also exist approaches that focus on Opinion Target Extraction to identify nouns that an opinion is describing. However, as described by Tomas Mikolov [4] and his team, tasks such as multi-class classification of words can be efficiently approached using word vectors that convert words into vectors of a predetermined length. Utilization of word vectors is a game changer when the applications of use are plenty. We found that this approach when paired with K-Means Clustering displayed the most accurate results.

3. Methodology

The system has been divided into 2 phases for ease of understanding:

1) Phase 1 - Dynamic Multi-Domain Scraping
2) Phase 2 - Review Trustability Score Calculation And Keyword Extraction

The model as seen in figure 1, represents complete system architecture comprising aforementioned phases.

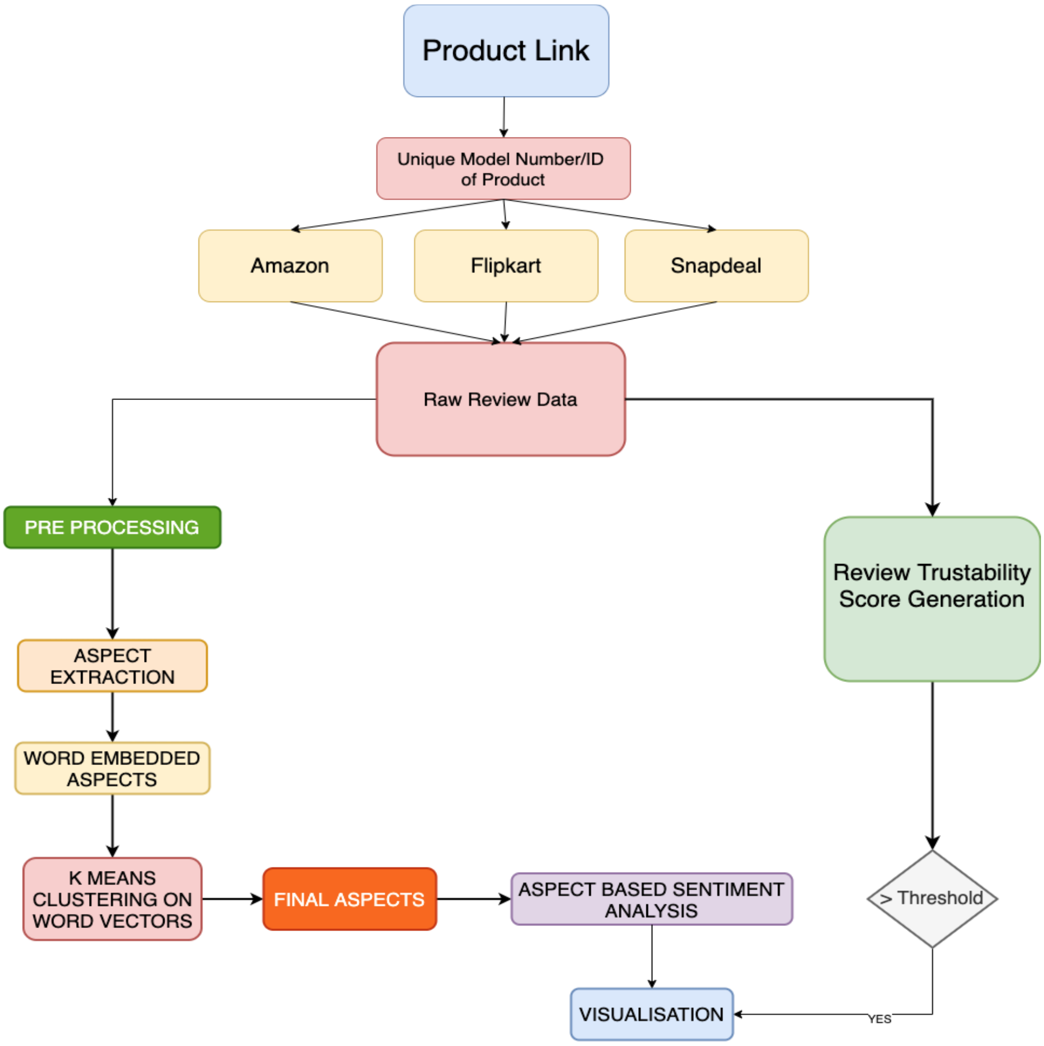


Figure 1. System Architecture

3.1. Dynamic Multi-Domain Review Scraping

The process of review scraping is performed using the python package Selenium. This package is used for automated browsing. Selenium let's one interact with webpages, browse different websites and parse HTML code. The method utilizes the unique model number of the product to search for it across various websites. It involves the following three steps.

1) ***Unique ID Identification:*** Using the link to the product, the system opens a browser and accesses the link using Selenium. From the webpage, the unique Model Number of the product is extracted by using its HTML tag ID. Using the Model Number, the product is searched on multiple domains. Currently it searches on Flipkart and Snapdeal.
2) ***Searching for the Product:*** The HTML code of the search result pages is parsed and the link of the respective products is used to identify the product. The review links of the identified products are parsed.

3) ***Scraping and Compiling Reviews:*** In the respective review pages, the reviews are identified by BeautifulSoup using the tag IDs and the text and details of the review are extracted. The data is stored as a pandas dataframe and saved to a .csv file.

As a result, we gather reviews of the product from multiple sources automatically. Each review contains the following attributes: Title, Rating, Description, Upvotes for the Review.

3.2. Review Trustability and Keyword Extraction

Review trustability and keyword extraction stages are formulated to identify trustable reviews and extract the features of the product discussed respectively. The extraction stage also covers end to end process steps.

3.2.1. Review Trustability Score: The trustability score of a review symbolizes how trustable it is, i.e how much the opinion can be valued. The score is based off of 4 key aspects of a review, they are:

i. ***Number of Sentences*** - Longer reviews are often more detailed and describe the product more accurately. Used the SpaCy package to count the sentences efficiently of every review.

ii. ***Readability*** - Readability is calculated by taking into account the kind of words used, sentence structures and other such characteristics. It describes how easy the text is to read. More readable reviews are generally supposed to be more trustable. We have used the Automated Readability Index algorithm to get the readability score for a particular review. This algorithm takes the word structure into account and is generally based on the average number of syllables per word or the proportion of easy words by referencing a word list.

iii. ***Target Words*** - Coming to the structure recognition of a review, a reader mainly focuses on certain words as he/she looks up at a review. We have taken the consideration of pros, cons, advantages and disadvantages under the target words and count the number of occurrences of these in each review.

iv. ***Number of Upvotes*** - Naturally the most upvoted reviews are more helpful.

Scores for each aspect are calculated and are used to determine the final trustability score for a review. The scores are normalized in the range [0,1]. Therefore, by taking the final scores of each review and checking if Score greater than Threshold (Average of Normalized Scores), we come to a conclusion which is represented at the visualization phase.
The weightage for the aforementioned aspects are as follows:

(V_1) Number Of Sentences	: 20% (0.2=W_1)
(V_2) Readability	: 30%(W_2)
(V_3) Target Words	: 20%(W_3)
(V_4)Number of Upvotes	: 30% (W_4)

$$\textit{Trustability Score} = N\left(\sum_{i=0}^{k} W_i * V_i \right) \qquad (1)$$

where,

N(x)- represents normalization function, resulting in the range of [0,1]
V_i - represents considered aspect's value

W_i - represents weightage given to the aspects based on the general behavior of the customer's mindset

3.2.2. Pre Processing: The review text that has been scraped from multiple domains has various unnecessary patterns, which is cleaned by the cleaning script we have formulated. It removes unnecessary characters, hyperlinks, symbols, excess spaces, and other text patterns that could not be processed by our algorithms. We have also replaced multiple period symbols with a single period. We extracted the text description for our analysis from this cleaned dataset. The other inconsistencies that prevail in the reviews scraped are the data type and data organization within their corresponding attributes. We have formulated mechanisms to typecast the data into required format and also trim the unnecessary data that comes along with the required data. The attributes such as votes are type casted into numeric format which is a primary requirement for further analysis.

3.2.3. Extraction of Noun-Adjective Pairs: The primary purpose of this step is to extract the aspects along with their corresponding modifiers (adjectives). We have used a python module named spaCy for the generation of dependency parse tree to extract aspect-adjective pairs based on specific syntactic dependency paths. The ultimate output of this step is a dictionary of such noun-adjectives that serve as input to the next grouping aspect step.

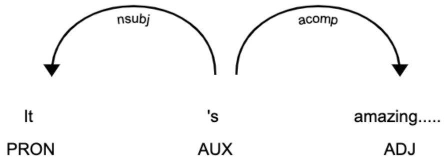


Figure 2. Dependency Parse Tree

The rules were formulated on the basis of the POS tagging of the words in the review text. For example, the noun of phrase would be a word with the "nsubj" dependence relationship with the verb token, and the adjective of that noun would be a word with the "acomp" dependence relationship. We also replaced product pronouns incurred in a phrase to 'Product', since the generic reference of the pronoun throws light onto the whole product. As a result, we would extract this pair as a relevant aspect-modifier pair. The image below illustrates some of the principles we formulated, where 'A' represents the aspect and M, M′ represents the respective modifiers of the aspect's adjective.

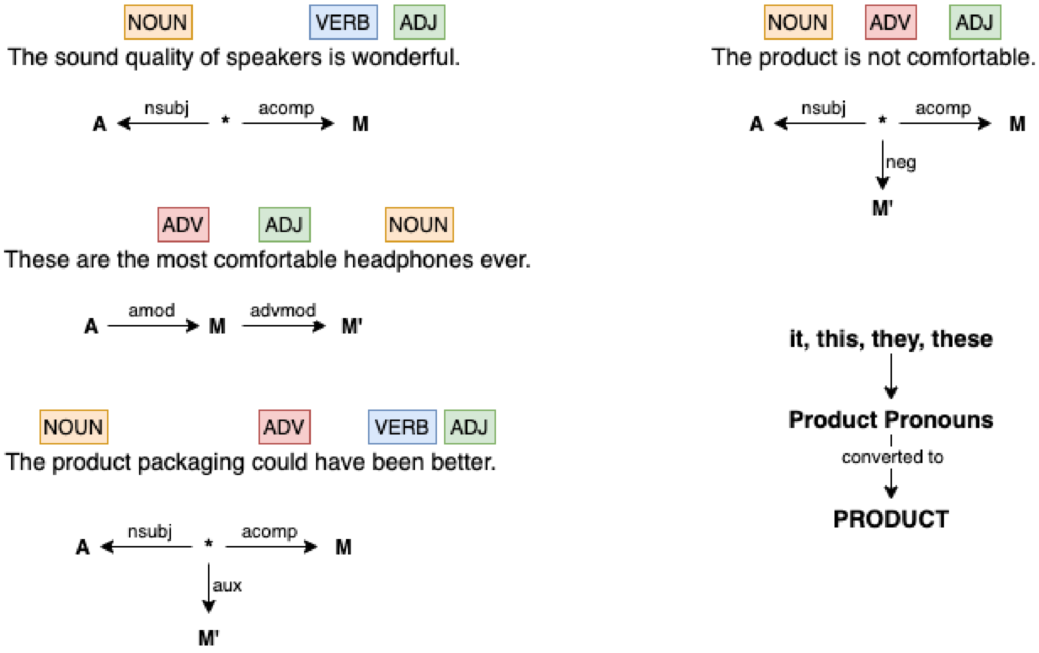

NOUN
VERB
ADJ
The sound quality of speakers is wonderful.
A
nsubj
*
acomp
M
ADV
ADJ
NOUN
These are the most comfortable headphones ever.
A
amod
M
advmod
M'
NOUN
ADV
VERB
ADJ
The product packaging could have been better.
A
nsubj
*
acomp
M
aux
M'
NOUN
ADV
ADJ
The product is not comfortable.
A
nsubj
*
acomp
M
neg
M'
it, this, they, these
Product Pronouns
converted to
PRODUCT

Figure 3. Rules Explanation

3.2.4. Grouping Aspects: This is done in two steps.

1) ***Word Embedded Aspects Generation*:** We establish the similarity of the word by contrasting the word vectors derived using spaCy's built-in high performance vectorisation model. Usage of spaCy for vectorization provides quick and easy access to over a million unique word vectors, and its multi-task CNN model is trained on 'web' data unlike 'newspaper' data as in other libraries like NLTK.
2) ***Clustering Aspects:*** The word vectors are then grouped using the K-Means algorithm provided by SciKit Learn. Varying its performance on the amount of data available, the K-Means algorithm gave us efficient results for 15 clusters. Clusters were labeled on the basis of the most frequently appearing word in each cluster.

3.2.5. Determining Polarity Scores: We used the Vader Sentiment Analysis method of the NLTK library, to determine the polarity of the aspect's adjectives. By noting the speed and accuracy as constraints, we chose this over spaCy and TextBlob methods. With the set of adjectives derived, we have calculated compound polarities of each of the adjectives and aggregated them in order to determine the final aspect polarity.

3.2.6. Visualization: We used the matplotlib library to visualize the analyses of review trustability and aspect based sentiment analysis(ABSA) of the product searched. Review trustability bar plots give an insight of the reviews analyzed and the demographics, whereas the ABSA bar plots give the final most discussed aspects throughout multiple domains of Flipkart, Amazon and Snapdeal with positive and negative bars of corresponding polarity values calculated.

4. Results and Discussions

The system that we built is able to dynamically extract reviews from multiple domains. Presently, the domains being used are Flipkart, Amazon and Snapdeal. Since Amazon and Flipkart are the most used websites, a large number of reviews were available for most products. Snapdeal consistently yielded less number of reviews. The process begins with the extraction of a unique model number from Flipkart which is further used to extract from Amazon and Snapdeal. After scraping the reviews, using the aforementioned techniques, the keywords of a text are extracted with corresponding adjectives, and the polarity for these aspects are calculated along with the trustability score for each review. The keywords or the aspects summarize the information in the text and provide a brief but powerful understanding of the content. Following is one such example on 'Dell Inspiron Laptop'.

4.1. Aspects

The system searches for the product from Amazon, Flipkart and Snapdeal, and presents the analyzed information in the form of bar plots. Lets observe the individual experiments conducted on Amazon, Flipkart and Snapdeal for product review analysis with our developed system and then compare with all of them together.

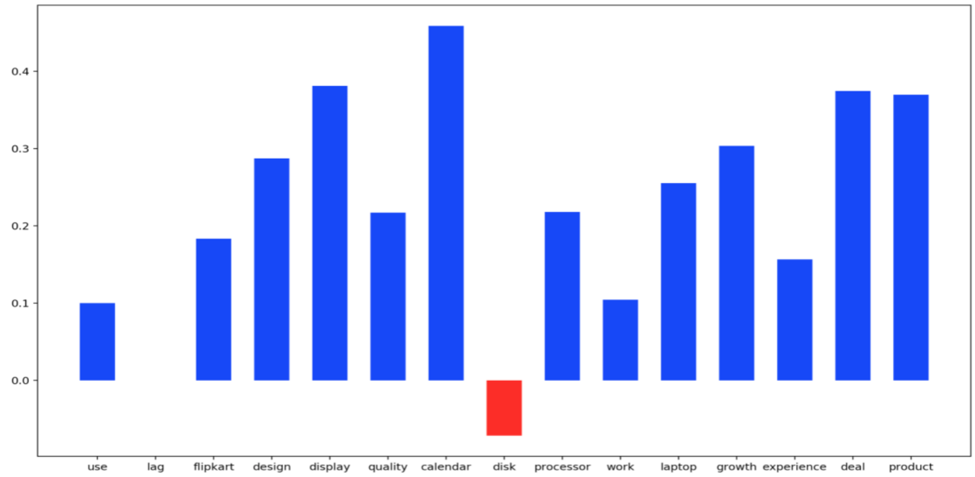


Figure 4. Amazon's Aspects With Corresponding Sentiments For Dell Inspiron Laptop

The above figure 4, represents the 15 most discussed aspects of the product across Amazon. We can observe the rich availability of information has enabled our system to capture a good amount of potential aspects that were discussed on Amazon. The aspects' opinions are mostly positive except a negative and a neutral opinion on a few aspects respectively.

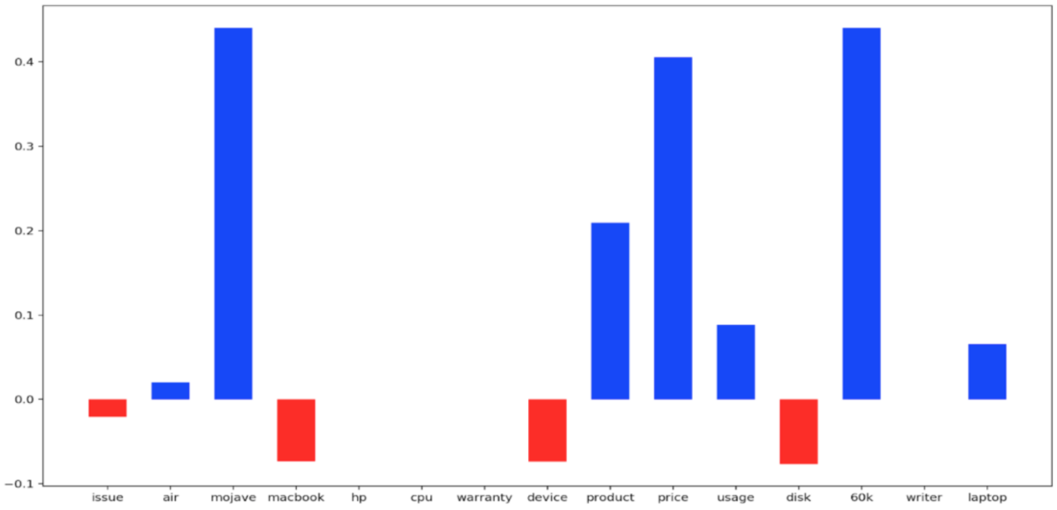


Figure 5. Flipkart's Aspects With Corresponding Sentiments For Dell Inspiron Laptop

The above figure 5, represents the 15 most discussed aspects of the product across Flipkart. The system was efficiently able to identify the aspects discussed which had a mixed behavior of the aspects discussed indicating positive, negative and neutral opinions.

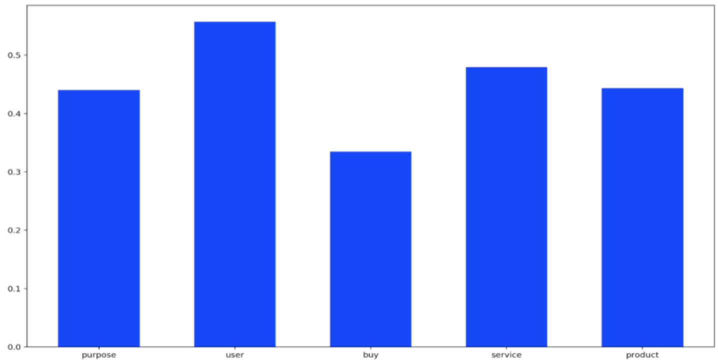


Figure 6. Snapdeal's Aspects With Corresponding Sentiments For Dell Inspiron Laptop

The above figure 6, represents the 5 most discussed aspects of the product across Snapdeal. The limited availability of information affects the analysis to put forth many aspects as seen in the before figures 4 and 5. This leads to the very known cold start problem due to limited data availability. Since the aspects are mostly positive unlike Amazon and Flipkart, a user is provided with ambiguous reviews and ratings. This is solved by our proposed system which considers the data of all the three aforementioned domains and analyses all together.

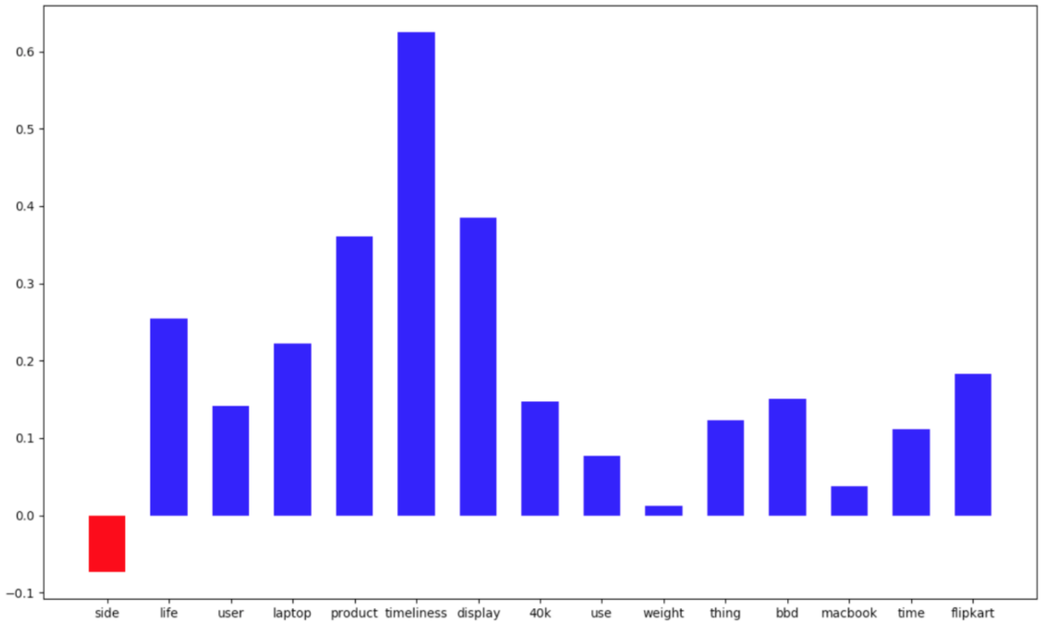


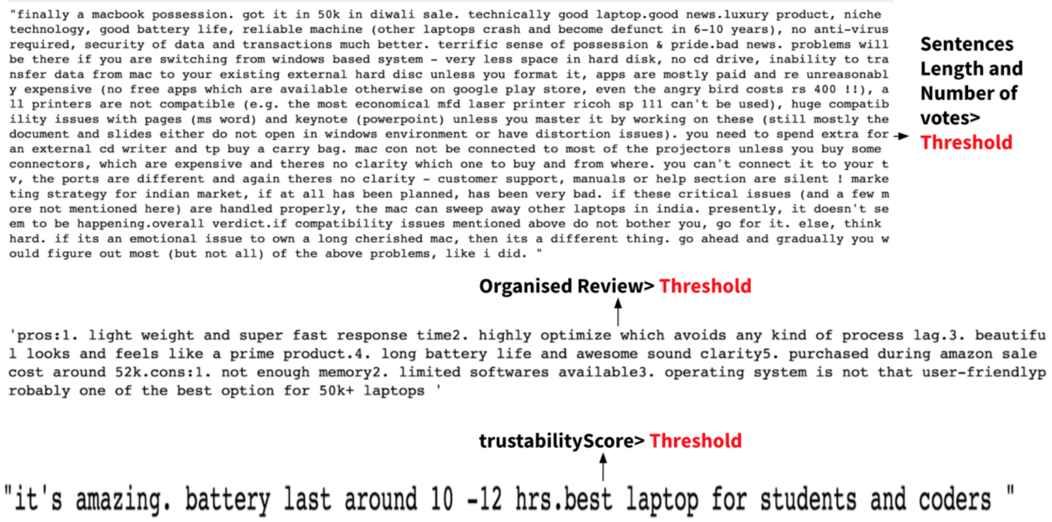
Figure 7. Dynamic Multi-Domain Aspects With Corresponding Sentiments

The above figure 7 represents the 15 most discussed aspects of the product across Amazon, Flipkart and Snapdeal. Unlike the existing methodologies, we cluster the aspects based on similarity of their meaning. Further, within each cluster, the most recurring aspect emerges as the final aspect of that cluster. This way, 15 clusters of different perspectives give 15 different aspects that efficiently represent the product. This prevents the bias in displaying top 15 recurring aspects discussed on the whole, but instead shows all the relevant aspects of all the perspectives discussed. This analysis stands ahead of all the other analyses by enhancing the richness in aspects and users' requirements.

4.2. Review Trustability

Trustability of a review is a crucial aspect solved in our research. We have allotted a trustability score for each review that is scraped from multiple domains by the aforementioned key features of length of sentences, readability of review text, number of upvotes and target words. By considering all these metrics, we overcome the bias that might arise by following the existing system's methodology of presenting the most helpful review solely with the number of upvotes.

In most e-commerce sites, the only quantitative metric for a customer to judge the trustability of a review is the number of helpful evaluations the review has. As seen in figure 8, our system considers 4 important measures of a review to assess its trustability(Votes, Sentences, Triggering Words and Readability). This process can tackle biases present in the existing system to guide a customer towards helpful reviews. A threshold is derived by the average of the normalized scores produced.



```
"finally a macbook possession. got it in 50k in diwali sale. technically good laptop.good news.luxury product, niche
technology, good battery life, reliable machine (other laptops crash and become defunct in 6-10 years), no anti-virus
required, security of data and transactions much better. terrific sense of possession & pride.bad news. problems will
be there if you are switching from windows based system - very less space in hard disk, no cd drive, inability to tra
nsfer data from mac to your existing external hard disc unless you format it, apps are mostly paid and re unreasonabl
y expensive (no free apps which are available otherwise on google play store, even the angry bird costs rs 400 !!), a
ll printers are not compatible (e.g. the most economical mfd laser printer ricoh sp 111 can't be used), huge compatib
ility issues with pages (ms word) and keynote (powerpoint) unless you master it by working on these (still mostly the
document and slides either do not open in windows environment or have distortion issues). you need to spend extra for
an external cd writer and tp buy a carry bag. mac con not be connected to most of the projectors unless you buy some
connectors, which are expensive and theres no clarity which one to buy and from where. you can't connect it to your t
v, the ports are different and again theres no clarity - customer support, manuals or help section are silent ! marke
ting strategy for indian market, if at all has been planned, has been very bad. if these critical issues (and a few m
ore not mentioned here) are handled properly, the mac can sweep away other laptops in india. presently, it doesn't se
em to be happening.overall verdict.if compatibility issues mentioned above do not bother you, go for it. else, think
hard. if its an emotional issue to own a long cherished mac, then its a different thing. go ahead and gradually you w
ould figure out most (but not all) of the above problems, like i did. "
```

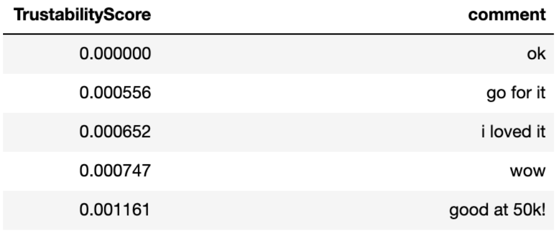
```
'pros:1. light weight and super fast response time2. highly optimize which avoids any kind of process lag.3. beautifu
l looks and feels like a prime product.4. long battery life and awesome sound clarity5. purchased during amazon sale
cost around 52k.cons:1. not enough memory2. limited softwares available3. operating system is not that user-friendlyp
robably one of the best option for 50k+ laptops '
```

```
"it's amazing. battery last around 10 -12 hrs.best laptop for students and coders "
```

Sentences Length and Number of votes> Threshold
Organised Review> Threshold
trustabilityScore> Threshold

Figure 8. Example of Reviews Recognised As Per Trustability Score Threshold

Hence, it can be concluded that the system built is an improvement over the review display system of e-commerce websites in most respects and offers the customer valuable information that he might not receive from browsing through an e-commerce website.



TrustabilityScore	comment
0.000000	ok
0.000556	go for it
0.000652	i loved it
0.000747	wow
0.001161	good at 50k!

Figure 9. Bottom Most Reviews Based On The Trustability Score

As seen in the figure 9, the reviews with mere description of a product aren't valued by customers and hence lower is the trustability score. This serves as a great asset for organizations as well, as it helps to extend the branch of use for feedback analysis as well.

5. Conclusion

The proposed work will help to eliminate the bias in the traditional commerce system to a large extent. As multiple domains are used as sources for extraction for a particular product, it helps increase the review corpus, which in turn gives a more accurate representation of the product, thereby reducing the users' complexity in operating various applications. It also helps unmask bias that is specific to a website. The system works well for products with a clearly defined and unique model number/ID. Hence, the system solves the cold start problem by dynamic multi domain scraping. One limitation of the scraping is the availability of data. Since the model number is identified for most important and tangible goods, the scraping is confined to those goods only. The aspect extraction phase uses natural language processing techniques of identifying noun-adjective pairs to identify aspects. This form of grammatical analysis is the most efficient way of identifying aspects. The clustering performed on the aspects assimilates knowledge of various aspects into a much smaller number of aspect clusters. It is important to note that the adjectives in questions are context-based. The resultant polarity values are based on these context-based adjectives and are more accurate than context-free sentiment analysis performed. The further enhancement of accuracy could be bolstered with the help of Recurrent Neural Network(RNN) [15], which handles the context of sentence structures that might not be identified by our proposed system. The trustability scores calculated are used as a threshold to filter reviews that are not considered trustable enough. The user can also consider the reviews with the highest trustability scores to get the most trustable opinion on the product. Since the least trustable scores are filtered out, the user gets a much more dependable analysis of the product and its features.

References

[1] SusantiGojali, Masayu Leyia Khodra, Aspect Based Sentiment Analysis for Review Rating Prediction, IEEE publication (2016).

[2] M. Hu and B. Liu, Mining and summarizing customer reviews, in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, (2004).

[3] A. M. Popescu and O. Etzioni, Extracting product features and opinions from reviews, in Natural Language Processing and Text Mining, Springer London, 2007, pp. 9-28.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean (2013). Distributed Representations of Words and Phrases and their Compositionality

[5] Araque, O., Corcuera-Platas, I., Sánchez-Rada, J., & Iglesias, C. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Systems with Applications, 77(19), 236–246.

[6] Basari, A., Hussin, B., Ananta, I., & Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. Procedia Engineering, 53, 453–462.

[7] X. Dai, M. Bikdash, B. Meyer, From social media to public health surveillance: Word embedding based clustering method for twitter classification, in: SoutheastCon 2017, 2017, pp. 1–7, http://dx.doi.org/10.1109/SECON. 2017.7925400.

[8] Q.V. Le, T. Mikolov, Distributed Representations of Sentences and Documents, Vol. 32,2014,http://dx.doi.org/10.1145/2740908.2742760.

[9] Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning.

[10] Giatsoglou, M., Vozalis, M., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. (2017). Sentiment analysis leveraging emotions and word embeddings.

[11] Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining (pp. 168–177).

[12] Iyyer, M., Manjunatha, V., Boyd-Graber, J., & Daume, H. (2015). Deep unordered composition rivals syntactic methods for text classification. In Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing: 1 (pp. 1681–1691).

[13] Kamkarhaghighi & Makrehchi, M. (2017). Content Tree Word Embedding for document representation. Expert Systems with Applications, 90, 241–249.

[14] Wieting, J., Bansal, M., Gimpel, K., & Livescu, K. (2016). CHARAGRAM: Embedding words and sentences via character n-grams. In Proceedings of the conference on empirical methods in natural language processing (pp. 1504–1515).

[15] LameiXu , Jin Lin, Lina Wang, Chunyong Yin, Jin Wang. Deep Convolutional Neural Network based Approach for Aspect based Sentiment Analysis,IEEE publication (2016).

Authors

Neelkanth Poosa is an undergraduate student in the department of IT in Chaitanya Bharathi Institute Of Technology. He has worked on projects in the domains of blockchain, web development and text analytics. His research areas include Natural Language Processing, web development, information security and block chain application development.

Pranav Sai Marla is an undergraduate student in the department of IT in Chaitanya Bharathi Institute Of Technology. He has worked on projects on the domains of sentiment analysis and artificial intelligence. His research areas include Natural Language Processing and Word Embedding.

Dr. M. Venu Gopalachari working as Associate Professor in Department of IT, Chaitanya Bharathi Institute Of Technology, Hyderabad. His research interests include data analytics, social media analytics and artificial intelligence.