International Journal of Engineering and Management Research
e-ISSN: 2250-0758 | p-ISSN: 2394-6962
https://ijemr.vandanapublications.com

Peer Reviewed & Refereed Journal
Volume-14, Issue-3 (June 2024)
https://doi.org/10.5281/zenodo.12541304

Sentiment Analysis of Social Media Data for Product and Brand Evaluation: A Data Mining Approach Unveiling Consumer Preferences, Trends, and Insights

Mahesh Prabu Arunachalam
Senior Manager, Department of Software Development and Engineering, Charles Schwab, Co, Texas, US

Corresponding Author: maheshprabu@gmail.com

Received: 11-05-2024 Revised: 28-05-2024 Accepted: 25-06-2024

ABSTRACT

Sentimental Analysis is an ongoing research field in Text Mining Arena to determine the situation of the market on particular entities such as Products, Services...Etc. This paper is a journal on sentiment analysis in social media that explores the methods, social media platforms used, and their application. It can be called a computational treatment of reviews, subjectivity, and sentiment. Social media contain a large amount of raw data that has been uploaded by users in the form of text, videos, photos, and audio. The data can be converted into valuable information by using sentiment analysis. We aim to collect details like Age, Gender, Education, Marital status, Salary, etc. So there requires data mining techniques like clustering. The Apriori Algorithm is the main algorithm used in our project. The Apriori algorithm is the general algorithm that can be used by developers according to their needs and implemented in their projects.

***Keywords*--** Sentimental Analysis, Big Data, Social-Media, Natural Language Process, Dataset, Apriori Algorithm, Association Rule, Regression Analysis, Orientation & Identification, Visualization

I. INTRODUCTION

The development of Web 2.0 is influencing the universe of virtual entertainment. Not just web-based virtual entertainment is used to associate and share data and their private belief with other people, however, even businesses can likewise impart, comprehend, and work on their items and administrations through interfacing in online entertainment. The quantity of web-based entertainment clients builds consistently and it is assessed in 2023 there will be 4.9 billion virtual entertainment clients around the world. Virtual entertainment is rich with crude and natural information and the improvement in innovation, particularly in AI and computerized reasoning, permit the information to be handled and changed into valuable information that can help most business association.

Wistful Examination is a computationally sorting and distinguishing sentiments communicated in a piece of the message which was to decide the client's, writer's, and essayist's surveys towards a specific item, topics, etc., separately and it tends to be likewise called a Message Mining. The significance of Message Examination and wistful investigation are compatible and the two words express shared importance, but a few specialists said that Assessment Mining and Nostalgic Investigation have unimportant various perspectives. Normal Language Handling (NLP) concentrates, converts, and understands assessments from a message and characterizes them into positive, negative, or regular opinions. The greater part of the past review applied feeling investigation into an item or film survey to all the more likely comprehend their client and settle on the important choice to work on their item or administration.

When they purchase the items online from internet business sites, will quite often rate these items and give audits on that item. This rating/survey framework frequently assists the other likely clients with choosing whether to buy that item or not. Nonetheless, perusing every one of the accessible surveys on a specific item, frequently causes the client to focus intensely on this interaction an overflow of spots, for example, web journals, audit destinations, and so on contain surveys. The course of feeling examination targets diminishing this season of the client by showing the information in a smaller configuration as means, investigation score, or basically histograms. The feeling examination methodology displayed in this paper can be reached out to the audits of items in various spaces. The trial results have shown that this strategy displays better execution. These days, individuals favor web-based shopping of items from different internet business sites since this assists them with saving time and offers them a more extensive scope of determination whenever the timing is ideal. Zeroing in our determination based on the buyer surveys of different clients assists us with saving time and channeling the items given the audits. Yet, a large portion of the surveys

Peer Reviewed & Refereed Journal
Volume-14, Issue-3 (June 2024)
https://doi.org/10.5281/zenodo.12541304

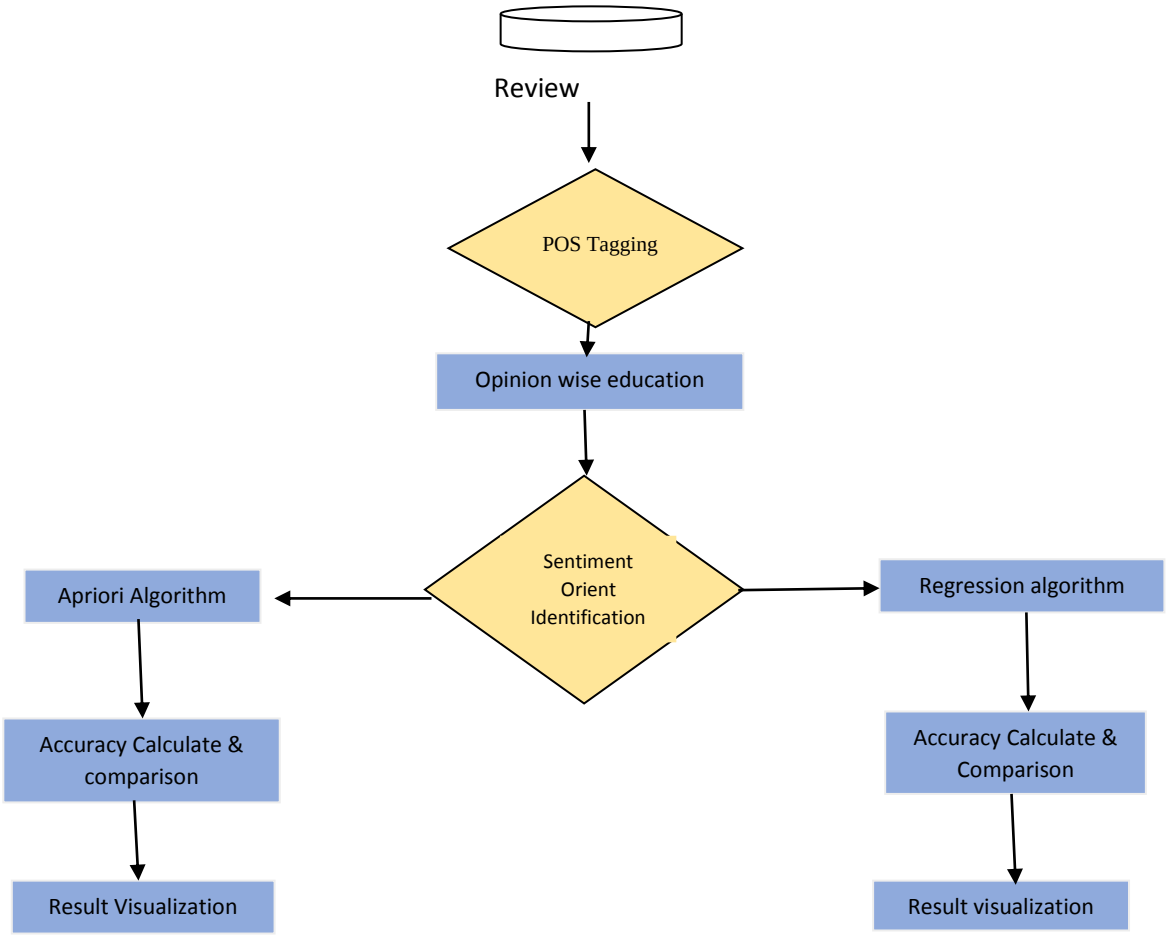
frequently contain exceptionally fewer insights concerning that specific item and have a greater amount of different sentences that are not helpful to the likely purchaser. Thus, we want to remove just that data expected by the client and trim out other undesired data, so they can be shown on reduced gadgets, for example, cell phones which are many times convenient among individuals. Along these lines, in this work feeling examination is anticipated for this reason. Feeling investigation is one of the phases of assessment mining. In opinion examination, we characterize the specific word as positive, pessimistic, or nonpartisan to anticipate the feeling of the speaker or analyst towards the item.

II. DATA MINING TOOLS

The initial form of data is the data present in a database. The first step of the data cleaning process is the removal of invalid data, the NULLs, and the transactions aborted mid-way. The next step of refinement is removal of the data which is not necessary for the analysis, If we are analyzing the products bought in India, there is no point in considering the data of the products bought in Germany. The next step is clustering, grouping of data which are similar. There are a bunch of parameters set for any client or a transaction, the transactions for which most or all of these properties match are grouped as similar. The result of such clustering is that when a new transaction or client registers with the same properties as set before, the output of the analysis can be utilized on that client. The final step in the process is transforming the data which we have refined till now into a format the pattern-searching algorithm expects.

III. METHODOLOGY

The modules of the system design are illustrated in the following diagram and explained in the subsequent sections.



Peer Reviewed & Refereed Journal
Volume-14, Issue-3 (June 2024)
https://doi.org/10.5281/zenodo.12541304

POS Tagging

Part-of-speech Tagging (POS Tagging) is the process of attaching every word of a file (corpus) with its corresponding part of speech, based on its definition and its relation with the adjacent phrases and words. The outcome of this process is all the words along with their equivalent POS tag from which the words can be identified as nouns, adjectives, pronouns, verbs, etc.

The process of POS tagging involves converting each word into Unicode Transformation Format (UTF-8) to encode all the character vectors into 8-bit code units to avoid complications of byte order marks. This is followed by tokenization of all the individual characters to convert them into individual tokens. The next step is to remove the stop words which are the most common words used in a language.

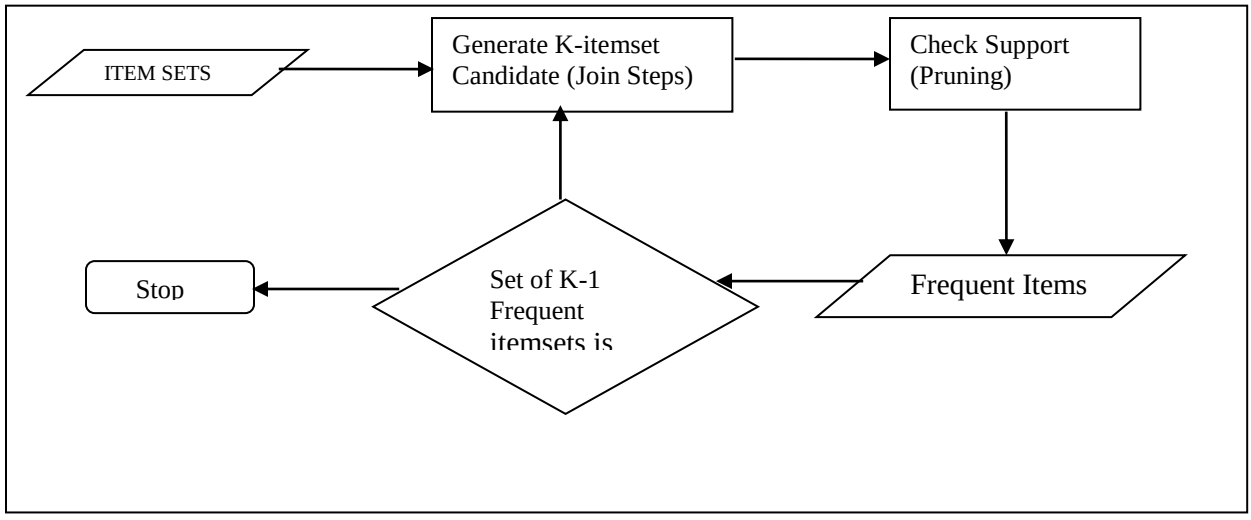
Apriori Algorithm

The process of POS tagging involves converting each word into Unicode Transformation Format (UTF-8) to encode all the character vectors into 8-bit code units to avoid complications of byte order marks. This is followed by tokenization of all the individual characters to convert them into individual tokens. The next step is to remove the stop words which are the most common words used in a language. The Apriori algorithm was given by R. Agrawal and R. Srikant in 1994 for finding frequent item sets in a dataset for the boolean association rule. The name of the algorithm is Apriori because it uses prior knowledge of frequent item set properties. We apply an iterative approach or level-wise search where k-frequent item sets are used to find k+1 item sets. This algorithm is also called the level-wise algorithm.

Steps in the Apriori Algorithm

The Apriori algorithm is a sequence of steps to be followed to find the most frequent items in the given database. This data mining technique follows the join and the prune steps iteratively until the most frequent items are achieved. A minimum support threshold is given in the problem or it is assumed by the user.

In the first iteration of the algorithm, each item is taken as a 1-itemsets candidate. The algorithm will count the occurrences of each item. Let there be some minimum support, minimum support. The set of 1 item set whose occurrence is satisfying the minimum sup are is determined. Only candidates that count as more than or equal to minimum support, are taken ahead for the next iteration, and the others are pruned. Next, a 2-item set of frequent items with min support is discovered. For this in the join step, the 2-item set is generated by forming a group of 2 by combining items with itself. The 2-item set candidates are pruned using the min-sup threshold value. Now the table will have 2 –item sets with min-sup only. The next iteration will form 3 –item sets using the join and prune step. This iteration will follow the antimonotone property where the subsets of 3-itemsets, that is the 2 –itemset subsets of each group fall in min support. If all 2-itemset subsets are frequent then the superset will be frequent otherwise it is pruned. The Next step will follow making 4-itemsets by joining 3-itemsets with itself and pruning if its subset does not meet the minimum support criteria. The algorithm is stopped when the most frequent item sets are achieved.



The Apriori Algorithm: Pseudo Code

- **Join Step:** C_k is generated by joining $L_{k-1}$with itself
- **Prune Step:** Any (k-1)-itemsets that are not frequent cannot be a subset of a frequent k-itemset

Peer Reviewed & Refereed Journal
Volume-14, Issue-3 (June 2024)
https://doi.org/10.5281/zenodo.12541304

- **Pseudo-code:** C_k: Candidate itemsets of size k
  L_k: Frequent itemsets of size k

L_2={frequent items};
for(k=1; t_k!= Ø; k++)do begin
C_{k+1}=candidates generated from L_k;
For each transaction t in the database do
increment the count of all candidates in C_{k+1} that are contained in t
L_{k+1} = candidates in C_{k+1} with min_support
end
return $U_k + L_k$;

Transaction Set for Apriori Algorithm

CUSTOMER_DETAILS

Email_Id	Name	Password	Gender	DateofBirth	MartialStatus	Education
ganesh1728@gmail.com	Ganesh	Ganesh	Male	28/02/1996	Single	Diploma
kajith@gmail.com	Ajith	Ajith	Male	1994	Single	Degree
manjunath@gmail.com	Manjunath	Manjunath	Male	12/09/1995	Married	PHD
sujayn@gmail.com	Sujay N	Sujay	Male	13/03/1995	Single	PU
ranjan@gmail.com	Ranjan	Ranjan	Male	1/1/1996	Single	Degree
ranjani@gmail.com	Ranjani	ABC	Male	1901	Single	Degree
sangeeta@gmail.com	Sangeeta	Sangeeta	Female	25/02/1995	Single	Master degree
sandesh@gmail.com	Sandesh	Sandesh	Male	31/3/1995	Single	Degree
spatika@gmail.com	Spatika	spatula	Female	23/04/1995	Married	10^{th}
swaroop@gmail.com	Swaroop	Swaroop	Male	12/04/1996	Single	Degree

ITEM_DETAILS

Item_ID	Sub-Category_ID	Item_Name	Item_cost	Item_Details	Quantity	attachment
1	1	Johnplayer	1000	Blue shirts	21	Null
2	2	Arrow	850	Casual shirt	28	Null
3	3	Nike	850	Sportswear	14	Null
4	7	Peter England	300	Classic Ties	42	Null
5	4	Bata	1500	Bata Shoes	29	Null
6	5	Django	1000	Casual Shoes	13	Null
7	6	Reebok	2000	Sports Shoes	21	Null

Peer Reviewed & Refereed Journal
Volume-14, Issue-3 (June 2024)
https://doi.org/10.5281/zenodo.12541304

INPUT DATASET

Trans-ID	Items
1	A, C, D
2	A, C, E
3	A, B, C, E
4	B, E

Minimum-Support=50%
Minimum-Confidence=80%
Item-set: A, B, C, D and E
STRONG ASSOCIATION RULE:
This is the result obtained.
1. {B}->{E}
2. {CE}-> {A}
3. {AE}->{C}
4. {A}-> {C}
5. {C}->{A}

The proposed system uses the apriori algorithm for finding item sets frequently bought together considering customer profile factors such as age, gender, income, education, and marital status rather than providing product-based recommendations. The customer spends more time due to better recommendations. This increases the profit for the organization and also increases the number of customers.

Regression Analysis

Regression analysis is a technique for determining numerically which elements might contribute. This technique is meant for companies that need in-depth, accurate, or quantitative knowledge of what might be impacting sales and how it can be modified, as necessary, in either direction.

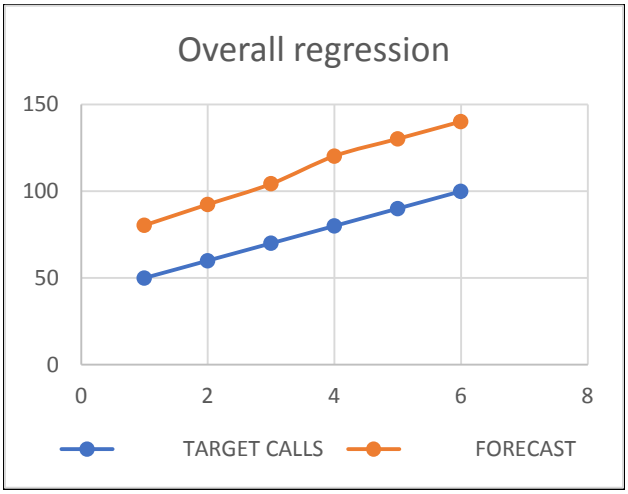
Data, which corresponds to the numbers and statistics that define your firm, is the key. The regression analysis has the benefit of enabling you to essentially crunch the data to aid in your choice for your company's present and future. Studying the relationships between data points is what the regression method of predicting implies, and this can help you with:

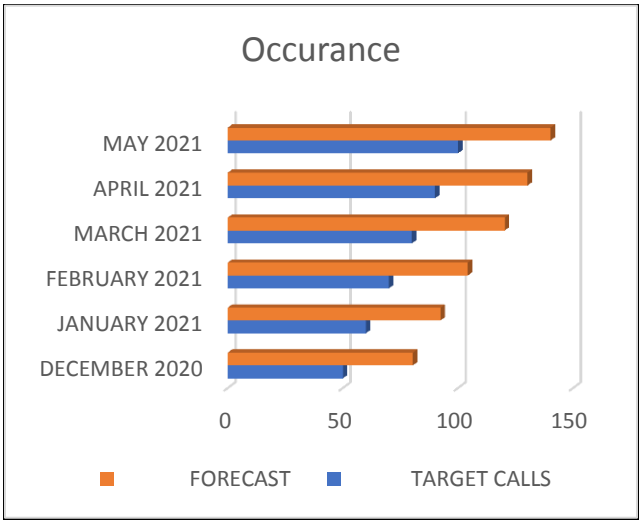
- Estimate both long- and short-term sales.
- Detect the inventory levels.
- Be conscious of supply and demand.
- Research and realize how different variables impact each of them.

"Since we're using Google Sheets, its built-in functions will do the math for us and we don't need to try and calculate the values of these variables. We simply need to use the historical data table and select the correct graph to represent our data."

Peer Reviewed & Refereed Journal
Volume-14, Issue-3 (June 2024)
https://doi.org/10.5281/zenodo.12541304

So, the overall regression equation is Y = bX + a, where:





To reiterate, I use the number 50 because I want to be sure that making more sales calls results in more closed deals and more revenue, not just a random occurrence. This is what the number of deals closed would be, not rounded up to exact decimal points.

SALES PERIOD	TARGET CALLS	FORECAST
DECEMBER 2020	50	80.38834951
JANUARY 2021	60	92.4173028
FEBRUARY 2021	70	104.27880279
MARCH 2021	80	120.2887727
APRIL 2021	90	130.210379
MAY 2021	100	140.1881572

Overall, the results of this linear regression analysis and expected forecast tell me that the number of sales calls is directly related to the number of deals closed per month. If you ask your salespeople to make ten more calls per month than the previous month, the number of deals closed will increase, which will help your business generate more revenue. While Google Sheets helped me do the math without any further calculations, other tools are available to streamline and simplify this process.

Sentiment Orientation Identification

The next step is to calculate the sentiment score of each review. The sentiment score helps us to classify the total score of each review and therefore the positive, negative, and neutral reviews can be identified. A score of +1 is assigned to a positive word whereas -1 is assigned to a negative word. The total review score can be calculated by summing up the individual scores of all the adjectives in a review. In this paper, the reviews with a score greater than 0 are classified as positive, reviews with a score of less than 0 are classified as negative, and a score of 0 makes the review a neutral review.

Result Visualization

The final result of classification can be represented in any format like bar graphs, histograms, trees, and tables. A histogram is used to show the results of sentiment classification. Receiver Operating Characteristic Plot (ROC Plot) to show the results of our analysis. An ROC Plot is a graphical plot depicting our classification results. The true positive rate is plotted against the false positive rate.

Peer Reviewed & Refereed Journal
Volume-14, Issue-3 (June 2024)
https://doi.org/10.5281/zenodo.12541304

IV. CONCLUSION

In conclusion, this paper highlights the significance of sentiment analysis in extracting valuable insights from social media data, particularly for evaluating products and brands. By leveraging advanced data mining techniques like the Apriori Algorithm, researchers can uncover consumer preferences, trends, and sentiments embedded within vast amounts of user-generated content. The study underscores the importance of understanding not only the sentiments expressed but also the demographic information of users to gain a comprehensive understanding of consumer behavior. By collecting details such as age, gender, education, marital status, and salary, researchers can tailor their analyses more effectively and derive actionable insights.

Furthermore, the paper emphasizes the versatility of the Apriori Algorithm, which serves as a powerful tool for identifying association rules within the data, thus enabling the discovery of meaningful patterns and relationships. Through the integration of natural language processing techniques and visualization methods, researchers can enhance the interpretability of results and facilitate informed decision-making processes for businesses. Overall, this research contributes to advancing the field of sentiment analysis by providing a comprehensive framework for extracting and analyzing sentiment from social media data. By leveraging big data analytics and computational methods, researchers can unlock valuable insights that can inform marketing strategies, product development, and brand management practices.

REFERENCES

[1] Statista. (2019) *Number of social media users worldwide 2010-2021*. Available at: https://www.statista.com/statistics/278414/number-ofworldwide-social-network-users.

[2] Giri, Kaiser J & Towseef A Lone. (2014). Big data-Overview and challenges. *International Journal of Advanced Research in Computer Science and Software Engineering, 4*(6).

[3] Sivarajah, Uthayasankar, Muhammad Mustafa Kamal, Zahir Irani & Vishanth Weerakkody. (2017) Critical analysis of big data challenges and analytical methods. *Journal of Business Research*.

[4] Agarwal, Basant, Namita Mittal, Pooja Bansal & Sonal Garg. (2015). Sentiment analysis using common-sense and context information. *Journal of Computational Intelligence and Neuroscience, 9*.

[5] U. T. Gursoy, D. Bulut & C. Yigit. (2017). Social media mining and sentiment analysis for brand management. *Global Journal of Emerging Trends in e-Business, Marketing and Consumer Psychology*.

[6] Devika MD, Sunitha C & Ganesh A. (2016). Sentimental analysis: A comparative study on different approaches. *Computer Science, 87*.

[7] Jandail RRS, Sharma P & Agrawal C. (2014). A survey on sentiment analysis and opinion mining: A need for an organization and requirement of a customer. *IJETAE, 4*(3).

[8] Rao NP, Nitin Srinivas S & Prashanth CM. (2015). Real time opinion mining of twitter data. *International Journal of Computer Science and Information Technologies, 6*(3).

[9] Singh T & Kumari M. (2016). Role of text pre-processing in twitter sentiment analysis. *Procedia Computer Science, 89*.

[10] Fillbrunn A, Dietz C, Pfeuffer J, Rahn R, Landrum GA & Berthold MR. (2017). KNIME for reproducible cross-domain analysis of life science data. *Journal of Biotechnology, 261*.