ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN2321-2152
www.ijmece .com
Vol 9, Issue.1Jan 2021

EFFICIENT PRODUCT REVIEWS APPROACH ON TWITTER USING DATA MINING

KALAKOTLA ARUNIMA,
Mr. E.BALAKRISHNA(Assistant Professor),

ABSTRACT

This paper aims to improve the methods and sequences for doing the necessary manipulation and extra prediction analysis. Our real-time dataset is derived from Twitter user comment sections.The unique thing about this dataset is that we have concentrated on certain comments that synthesize a phrase using the product. The dataset is then filled up by doing more iterations of extraction. Our three-column dataset is based on the assumption that the subject or user has enhanced some part of the product being studied. We set out to investigate the actual use of products manufactured by IT behemoths like Apple and Google using this particular dataset. With their next cyber projects and more advanced items, the two tech giants want to enhance their technical domination, which they are already acutely aware of.Reason being, they are always in the process of making their devices better and adding more features to them.It will be difficult for them to finish their present undertaking without the feedback, advantages, and disadvantages of their earlier efforts.In order to further enhance their product, they may extract these properties from the dataset provided by the Twitter API. They could look at social media to find out whether the product has been successful after it hits the market, evaluate the downsides, and avoid taking a poll altogether.Such a classification system would be more tedious in the long run and would not get the results we need. The next step is to start mining social media for data.

I. INTRODUCTION

A user's information is a crucial component in determining how to classify that user's characteristics. One of the most important means of disseminating news and information nowadays is social media [1]. The transmission of vital information from one end of the world to another is the primary goal of internet development. The transmission of data allows for the distinctive enhancement of information and its eventual access to the people. On the other hand, advertising products are supplied by marketing and firm companies that utilize the data. Additional machine learning techniques[2] were used and modified inside the users' dataset. To begin, either the dataset belonging to a specific user is extracted or several datasets belonging to a massive number of users are extracted. After that, the data sets will need to be extracted at regular intervals. Unstructured data would make up the framework[3] in the outset. It is possible to decompose the data into chronological order after receiving the structured data. The last step is to utilize machine learning and statistical analysis on the retrieved structured data on a user's

Department:CSE,
Vaagdevi College of Engineering ,Address: Bollikunta, Warangal Telangana,
Pincode:506005.

logins, logouts, picture posts, and tagging to derive a personality assessment. Concerning privacy, the "Research Challenge on Opinion Mining and Sentiment Analysis" provides a conceptual view. A little amount of theory and user psychology on social media trafficking has been culled from this article. While the technique remains unchanged, we have implemented a real-time model that supplements it with theoretical analysis. We have also included the backend method for how user privacy is utilized as data, in addition to the theory-based principles. This study primarily presents ideas on Facebook manipulation and the ways in which it influences users' access. The major conceptual paper that we have taken for our reference purpose clearly explains its users' experience.

II. LITERATURE SURVEY

When creating software, the first and first thing to do is a literature review. It is important to assess the time factor, economic situation, and corporate strength before to building the tool. After these prerequisites are met, the tenth step is to choose an OS and programming language that may be used to create the tool. As soon as the developers begin working on the tool, they will need substantial assistance from other sources [6]. Help with this may be found online, in books, or from more seasoned programmers. All of the aforementioned factors are considered while designing the system before construction begins.

Research challenge on Opinion Mining and Sentiment Analysis[7]

AUTHORS: David Osimo and Francesco Mureddu

Opinion mining and sentiment analysis are the focus of this study, which aims to lay down the groundwork for a new research challenge in these areas. By defining a new Research Roadmap on ICT Tools for Governance and Policy Making, this research challenge has been developed within the scope of project CROSSOVER "Bridging Communities for Next Generation Policy-Making." It builds on the model and research roadmap developed within the scope of the CROSSROAD project, but with a stronger focus on governance and policy modeling. With this goal in mind, CROSSOVER is concentrating on modifying two Grand Challenges—GC1 - Model-based Collaborative Governance and GC2 - Data-powered Collective Intelligence and Action—that are currently included in the CROSSROAD roadmap. Research challenges are a component of each Grand Challenge. Specifically, the research challenge "Peer to peer public opinion mining" is embedded in Grand Challenge 2, and our goal for the workshop is to revise, update, enhance, and verify it.

Stalker, A Multilanguage platform for Open Source Intelligence[8]

AUTHORS:Neri F, Massimo Pettoni

Thanks to the IT revolution, open sources are becoming more useful, widely used, and accessible than ever before. In recent years, open sources have become more accessible and affordable for the worldwide Intelligence Communities. However, the vast majority of electronic data consists of text, and the most important details are often concealed and encoded in unclassified or poorly organized pages. The capacity to manage the challenges of multilingualism is closely tied to automated textual analysis and synthesis, which is essential for obtaining and processing this vast amount of raw material that is linguistically and source varied. Experts and the general public alike may benefit from the content enabling system detailed in this article, which allows for deep semantic search and information access to massive amounts of distributed multimedia data. For a wide variety of data

gathered from several sources in a variety of culturally distinct languages, STALKER offers language independent search capabilities and dynamic categorization features.

Mining Textual Data to boost Information Access in OSINT

AUTHORS: Neri F, Geraci P

Thanks to the IT revolution, open sources are becoming more useful, widely used, and accessible than ever before. Access to open source information has been steadily improving and becoming more affordable for the world's intelligence communities in the last few years. The majority of electronic data consists of text, with the majority of useful information stored on unstructured and unclassified pages. The capacity to manage the challenges of multilingualism is crucial to the process of accessing and changing all this raw data, which is diverse in terms of the languages employed. This in turn is connected to the notions of textual analysis and synthesis. In order to facilitate the cycle of information gathering, processing, exploitation, creation, distribution, and assessment, some Intelligence operating structures in Italy have used SYNTHEMA SPY Watch, a content enabling system for OSINT. The technology provides operational officers with a high-level view of massive amounts of textual data, allowing them to uncover all relevant information and significant commonalities between documents.

A Multilanguage platform for Open Source Intelligence[10]

Baldini N, Neri F, Pattoni M.

The term "Open Source Intelligence" (OSINT) refers to a method of gathering intelligence that makes use of data retrieved from publicly available sources. Open intelligence is now more affordable than ever before because to the information technology revolution, which is also making open sources more accessible, pervasive, and useful. Human intelligence (HUMINT), intelligence on aerial imagery (IMINT), and signals intelligence (SIGINT) are all seeing open versions brought about by the growth of open source intelligence (OSINT), which is revolutionizing the intelligence community. In recent years, open sources have become more accessible and affordable for the worldwide Intelligence Communities. However, the vast majority of electronic data consists of text, and the most important details are often concealed and encoded in unclassified or poorly organized pages. The capacity to manage the challenges of multilingualism is closely tied to automated textual analysis and synthesis, which is essential for obtaining and processing this vast amount of raw material that is linguistically and source varied. In order to handle massive datasets gathered from diverse and geographically dispersed information sources, this study details a multilingual indexing, searching, and clustering system that offers language independent search capabilities and dynamic classification capabilities. This system has been implemented by the Joint Intelligence and EW Training Centre (CIFIGE), a military school, to educate both military and civilian workers of the Defense Department in the field of OSINT.

[9]

III. EXISTING SYSTEM[11]

While some social media platforms take extra precautions to protect their users' personal information, others make it publicly available, making it easy for anybody to manipulate. Users' privacy agreement status ensures that Facebook safeguards their data, however some companies pay a fair price to get the Facebook domain and utilize it to get users' data. Due to the large number of users'

ISSN2321-2152
www.ijmece .com
Vol 9, Issue.1Jan 2021

private messages stored in the data vault, Facebook security is top-notch. However, they are able to sell the data beyond this. However, data scientists are able to build projects and do exact analyses since Twitter generates comments and makes the data publicly available. Every user's remark is randomly selected from the comments made on the iPad using the Google service. Finding the reach rate of the following goods using Twitter user comments is the major purpose of this study.

Disadvantages of Existing System[12]

1)Completing the paper without feedback would be a challenging undertaking. 2)Instead of doing a poll, all of this information may be retrieved via social media now that the product has successfully entered the market.We can't anticipate reliable outcomes from such a method of categorization, and it would be more laborious overall.

IV. PROPOSED SYSTEM

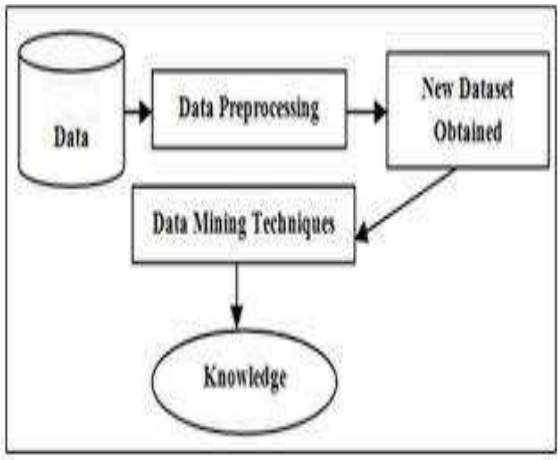

Fig 1. System Architecture

Using a two-stage suggestion technique, this study tackles the difficulty of helping consumers write higher-quality product review Tweets: terms for Product Features: Suggests certain terms for product features. Appropriate Opinion Word Recommendation: Suggests suitable opinion words to characterize the associated product attributes.

3.1 Recommending Product Feature Words

Assuming that a single phrase may adequately represent a particular feature or subject, reviews are often divided down into sentences. In order to identify the different parts of speech in the review, the phrases have been Parts-of-Speech (POS) tagged. Then, we determine that nouns are the most common words to describe the product's attributes. We are just concerned with feature nouns, and the difficulty comes from trying to differentiate them from non-feature nouns. Since the consumers are interested in expressing their thoughts regarding feature nouns, we continue with our experiment assuming that they appear next to adjectives. Contrarily, non-feature nouns do not fall under this category.

See, friend, mobile, and features are the nouns that pop up in the statement. The terms "look," "mobile," and "features" are most relevant to our discussion since they all pertain to mobile phones. The users want to convey their ideas about the characteristics, thus the nouns mobile, appearance, and features all have adjectives linked with them. However, the term buddy does not. The user has no interest in sharing their opinion on it in a review post as it is unrelated to mobile phones. Using this finding, we can sort potential feature noun candidates from those that aren't. After keeping track of where the feature noun candidates are in the phrase, the following steps are taken to prepare the file for the LDA algorithm. One useful use of LDA is the extraction of latent themes from datasets. A topic is really just a cluster of words that

ISSN2321-2152
www.ijmece .com
Vol 9, Issue.1Jan 2021

are somehow connected to one another. A review dataset's subject is a collection of feature words that are connected to one another. Flash, pixel, front, back, digital, wide, and so on are all examples of feature words that may be grouped together under a single, latent subject camera. To make the topic's Tweet more informative, you may utilize these relevant feature terms. As a result, the user is presented with suggestions based on the extracted significant elements of any given subject whenever they begin to compose a Tweet on that topic. This feature enhances the informativeness of the Tweet. The first algorithm provides a pseudocode summary of the procedure.

Input: Product reviews from the corpus
Output: Product features and feature topics

```
for each review R in the corpus do
    for each sentence S in R do
        for each word W in S do
            POS_Tag(W)
            if POS_Tag(W) == NN then
                if POS_Tag(W) preceded by JJ then
                    Retain W in S as Candidate Feature Noun
                else
                    Delete W
Product features and topics= LDA (Review Sentences with
    Candidate Feature Nouns)
```

Algorithm 1: Identify Product features and feature topics

Recommend Appropriate Opinion Words

Using the POS Tagged reviews as a starting point, we extract every adjective from the dataset. Along with the review text, each review in the review dataset also contains a numerical rating. There is a star rating system that goes from 1 to 5. In order to categorize the reviews, we look at the star rating.

We use the adjectives discovered by POS tagging and count how often they appear in each of the five review categories to determine the sentiment words' polarity. The second algorithm delves further into the process of sentiment word categorization using star ratings.

Input: POS Tagged Review text, star rating associated with the reviews. **Output:** Sentiment words and associated polarity.

```
for each review R in the POS tagged Reviews do
    for each sentence S in R do
        for each word W in S do
            if POS_Tag(W) == JJ then
                Add W to the Adjectives_List
for each review R in the corpus do
    if R has 1 star rating then
        Classify R as 1 star review.
    else if R has 2 star rating then
        R as 2 star review
    else if R has 3 star rating then
        Classify R as 3 star review.
    else if R has 4 star rating then
        Classify R as 4 star review
    else if R has 5 star rating then
        Classify R as 5 star review
    else
        Discard R
for Adjective A from Adjectives_List do
    if if A occurs majority times in 4 or 5 star reviews then
        Label A as positive
    else if if A occurs majority times in 1 or 2 star reviews
    then
        Label A as negative
    else
        Label A is neutral
```

ISSN2321-2152
www.ijmece .com
Vol 9, Issue.1Jan 2021

Algorithm 2 Identify the domain based polarity of sentiment words

Then, for each attribute, we get the frequency of all emotion words: positive, negative, and neutral.
We determine the TF-IDF of sentiment adjectives in the positive, negative, and neutral sentiment categories with regard to each feature in order to identify feature specific adjectives.
The most intense emotion word for that characteristic is the sentiment adjective with the greatest TF-IDF.

For each set of characteristics, we keep the 25 adjectives that scored the best in the positive, negative, and neutral emotion categories according to the TF-IDF. Users will be able to see these generated opinion words as potential suggestions when they express their thoughts on certain features in Tweets. The third algorithm provides more detail using pseudocode.

Input: 1. Product Features Extracted using Algorithm 1 denoted as F,
2. Set of Sentiment categories: Positive, Negative, Neutral Extracted using Algorithm 2 referred as S
Output: Appropriate sentiment words with intensities for features identifies.

1: **for** Feature F_i in F **do**
2: **for** Sentiment Category S_i in S **do**
3: **for** Sentiment Word SW_i in S_i **do**
4:

$$TF(SW_{iFi}) = \frac{\text{Frequency}(SW_i)_{Fi}}{\sum_{i=1}^{n}(SW_i)_{Fi}}$$

5:

$$IDF(SW_{iFi}) = \log_e \frac{\text{Number of features}}{\text{Features with} SW_i}$$

6:

$$TF - IDF(SW_{iFi}) = F(SW_{iFi}) * IDF(SW_{Fi})$$

7: **for** Feature Fi in F **do**
8: **for** Sentiment Category Si in S **do**
9: Sort SW based on $TF - IDF(SW_{iFi})$

Algorithm 3 Generating feature based sentiment words recommendations

The primary goal of this study is to make predictions on the intentions of the observed product's users. The modification might be done using a variety of machine learning approaches, but we're taking a more targeted approach. Consequently, our product's regression analysis will make use of a single categorization approach. One possible outcome of our assignment is to forecast the product's utilization. Based on the comment sections of Twitter users, we have obtained a real-time dataset. What makes this dataset special is that we have zeroed in on specific remarks that use the product to synthesize a term. The dataset is then filled up by doing more repeated extraction. Based on the belief that the user or subject has improved any aspect of the product under observation, our dataset comprises three columns. With this specific dataset, we set out to examine how people really utilize items made by tech giants like Google and Apple. Both companies are well-versed in the technological period in which they now rule, and they are planning to further concentrate on cyber projects in the future, resulting in more sophisticated items.Because they are actively working to improve their gadgets and make them more feature-rich.Without the comments, benefits, and drawbacks of their previous initiatives, it would be difficult for them to complete their current endeavor.
They have the ability to extract these attributes from the Twitter API dataset, which will allow them to develop their offering even more. In place of a poll, they may look to social media to see whether the product has been successful in the market and if there are any downsides.We can't anticipate reliable outcomes from such a method of categorization, and it would be more laborious overall. We will now begin

ISSN2321-2152
www.ijmece .com
Vol 9, Issue.1Jan 2021

the process of data mining from social media.

Advantages of Proposed System

1) The proposed system's strength lies in its ability to identify and guard against newly-discovered malware, even with a 2-10% variance. Machine learning-based system security applications are able to decipher the pattern of code.

V RESULTS

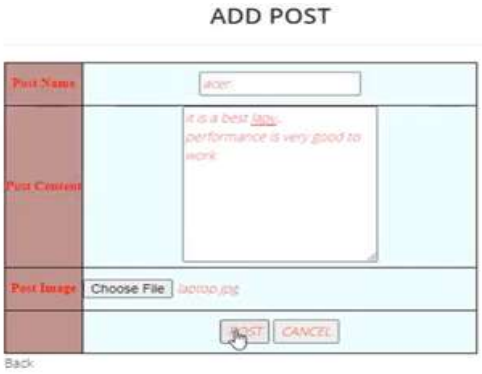

Fig 2. How to Add Post

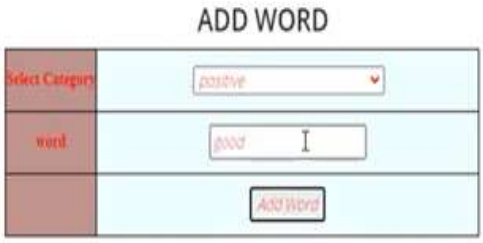

Fig 3. How to Add word

Fig 4. How to view Post

Fig 5. How to add comment

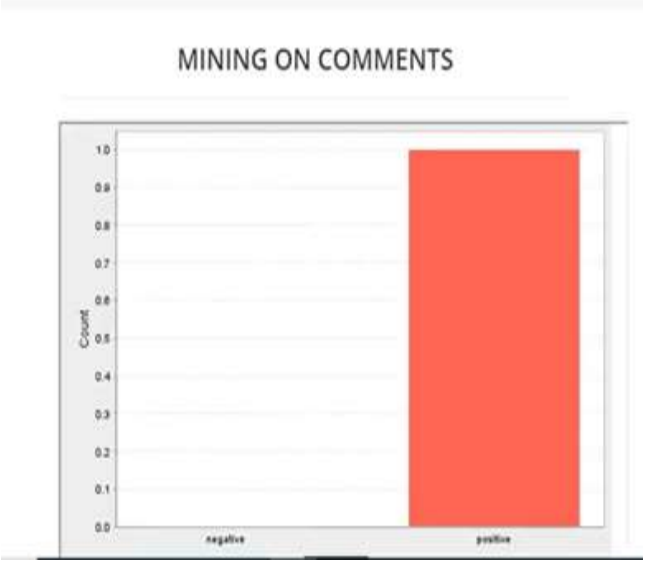

Fig 6. Mining On Comments

CONCLUSION

Many methods exist for email clients to detect and block spam. Machine learning ensures that these spam filters are constantly updated. It is impossible to keep up with the newest spammers' techniques while using rule-based spam filtering. When it comes to spam filtering, ML is the engine that drives approaches like Multi Layer Perception and C 4.5 Decision Tree Induction. On a daily basis, more than 325,000 malwares are identified, with each piece of code being 90-98% identical to its earlier iterations. Machine learning-based system security applications are able to decipher the pattern of code. Consequently, they are able to swiftly identify and provide protection against new viruses with a variance of 2-

10%. An increasing number of websites now let users talk with a support agent without ever leaving the site. Having said that, not all websites have a real executive to respond to your questions. Typically, you will be interacting with a chatbot. These automated programs often scrape data from websites and display it to users. As time goes on, the chatbots become smarter. The machine learning algorithms allow them to better comprehend user inquiries and provide better replies. Consequently, we may improve each real-time application with the greatest accuracy rate by thinking about the best categorization approach.

Future Enhancement

We can improve each real-time application with the greatest accuracy rate by evaluating the best categorization approach.

I will enhance the accuracy.

Cut Down on Runtime

Shorten the time it takes to detect.

REFERENCES

1. David Osimo and Francesco Mureddu, "Research challenge on Opinion Mining and Sentiment Analysis"

2. Maura Conway, Lisa McInerney, Neil O'Hare, Alan F. Smeaton, Adam Berminghan, "Combining Social Network Analysis and Sentiment to Explore the Potential for Online Radicalisation, Centre for Sensor Web Technologies and School of Law and Government.

3. A.Cutillo,R.Molva and T.Strufe.Safebook:A privacy preserving online social network Leveraging on real life.

4. S.Buchegger,D.Schioberg,L.Vu p2p social networking –early experience and insightsSNS2009.

5. Lucas C., "Sentiment Analysis a Multimodal Approach," Department of Computing, Imperial College London_, September 2011.Celli, F., Pianesi, F., Stillwell, D. S., and Kosinski, M. 2013.Workshop on Computational Personality Recognition

6. Mining Facebook Data for Predictive Personality ModelingDejanMarkovikj, Sonja Gievska ,MichalKosinski, David Stillwell

7. Center of Attention: How Facebook Users Allocate Attention across Friends Lars Backstrom, EytanBakshy, Jon Kleinberg ,Thomas M. Lento, ItamarRosenn

8. Sentiment Analysis for Social Media R. A. S. C. Jayasanka, M. D. T. Madhushani, E. R. Marcus, I. A. A. U.Aberathne ,and S. C. Premaratne

9. Pang, B., Lee, L., Vaithyanathan, S.: "Thumbs up? sentiment classification using machine learning techniques", in Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Volume 10, July 2002, pp. 79-86.

10. Socher, R., et al.: "Semi-supervised recursive autoencoders for predicting sentiment distributions" in Proceedings of EMNLP '11 - the Conference on Empirical Methods in Natural Language Processing, ISBN: 978-1-937284-11-4, pp. 151-161

11. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: "Lexicon-Based Methods for Sentiment Analysis", in "Computational Linguistics", June 2011, Vol. 37, No. 2, pp. 267-307.

12. Chaovalit, P., Zhou, L.:. "Movie review mining: A comparison between supervised and unsupervised classification approaches", in Proceedings of the Hawaii International Conference on System Sciences (HICSS), 2005.