IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Contextual and Omni Channel Sentimental Analysis

[1]P.Karthikeyan, [2]P.Aruneshwar, [2]S.Deepak, [2]R.Vignesh

[1]Professor, Department of Computer Science and Engineering, Sri Manakula Vinayagar Engineering College, Puducherry, India.
[2]UG Student, Department of Computer Science and Engineering, Sri Manakula Vinayagar Engineering College, Puducherry, India.

Abstract: The internet is one of the rapidly evolving areas that is transforming people's lifestyles. They frequently use the internet to communicate, discuss, and share information. As a result of these factors, the internet has become an integral aspect of human life. It contains information on a wide range of topics, including academic knowledge, product feedback or opinions, comments on social issues, and much more. It aids people in thinking and making decisions in a variety of situations. Before making a final decision, the majority of people seek advice from others. Sentimental analysis is one of the research areas in which data is acquired and evaluated in order to determine the sentiment of the data, such as negative or positive sentiment. This sentimental analysis, when applied to online-sold products, will benefit both customers and business entities, as buyers will be able to gain a sense of the product and business entities will be able to use the obtained data to improve the product. Existing sentimental analysis systems are unable to accurately classify sentences that contain negative polarity words but do not convey the same emotion, and these models ignore modern generation jargon like FOMO, TTYL, and others, which may provide contextual meaning and, if ignored, may cause a deviation in the actual sentiment conveyed by the sentence. To overcome the above-mentioned drawbacks, our system can analyse modern world jargon, and the actual context of the sentences is obtained through an algorithm that uses contextual analysis, diagram method, and ensemble feature selection, and the results can be used for data analysis and data visualization, which can then be used to improve the products.

***Index Terms* – Sentimental Analysis, contextual analysis, polarity, ensemble feature selection, data visualization.**

I. INTRODUCTION

People are more aware of what is going on around them as a result of the emergence of social media and e-commerce sites, which has resulted in a rapid expansion of data. Information and knowledge finding that is both effective and instantaneous has become a vital part of daily life. As the demand for information grows, the attention is shifting to analysis methods. People are curious about other people's viewpoints. Consumers are encouraged to leave reviews on Most retail websites to convey their thoughts on various elements of the merchandise. Consumers frequently utilize internet reviews to gather useful information before purchasing a product, and many businesses rely on them for product development, marketing, and customer relationship management or CRM. Some characteristics, we suggest, are more significant than others, and have a stronger impact on customers' final decisions and enterprises' product development strategies. As a result, identifying key product features will increase the usability of various reviews, which will benefit both customers and businesses. Consumers may make better informed purchasing decisions by focusing on the important factors, while businesses can focus on enhancing the quality of these aspects and therefore effectively boost product reputation.

II. LITERATURE SURVEY AND RELATED WORKS

Deepanshi and Adwitiya Sinha gathered real-time user comments and feedback from two meal delivery firms' Twitter feeds. Following that, natural language analytics is used to extract the most common circumstances. The proposed algorithmic methodology is also utilized to create a signed social network in order to analyse product-centric behavioural sentiment. Sentiment analysis at the fine-grained level concerning settings provided a larger perspective for evaluating and performing contextual predictions. Customer behaviour is studied, and the results are interpreted in both positive and negative light. The social behavioural model's findings predicted customers' positive and negative contextual feelings, which may be utilized to help determine future initiatives and ensure service quality for improved customer satisfaction. Not only does author K Barnes provide a window into Internet users' thinking during the COVID-19 outbreak, but he also conducts a content-based prediction analysis of what makes a meme go viral. We also investigate the incremental predictive potential of image-related characteristics over textual features on meme popularity using machine learning approaches. We discovered that the popularity of a meme can be predicted relatively well based on its content alone; our best machine learning model predicts viral memes with an AUC of 0.68. Furthermore, we also discovered that picture and textual features have significant incremental predictive value over one another. They suggested a new paradigm for multi-domain active learning. This framework selects text data from Amazon.com across all domains, including BOOKS, DVDs, Electronics, and Kitchen. The

Multi-Domain Sentiment Dataset is the data collection used by this system. For weighting features, the authors of this work used phrase frequency. They employed LIBLINEAR SVM to construct a superior classification model. Po-iet presented a novel method for extracting feelings from microblogs, which he called Po-iet. They discovered that while some tweets are favourable, when it comes to sentiment, they are negative. Unigram was the feature employed in this project. They employed mutual information and chi-square as feature selection approaches to limit the number of features in the collection. The lexicon-enhanced technique was used to create a collection of sentiment words based on the sentiment lexicon. These words have been used as a novel feature in this publication. Sentiment terms, as well as content-specific and content-free aspects, were used in this investigation. The evaluation was carried out with the use of 10-fold cross validation. Hesham Arafat's findings reveal that for sentiment classification, mRMR outperforms IG, and that a hybrid feature selection method based on RST and Information Gain (IG) outperforms the prior methods. The proposed approaches are tested on four typical datasets: movie reviews, product reviews (books, DVDs, and gadgets), and sentimental categorization. Experimental findings suggest that the hybrid feature selection method outperforms feature selection methods. Asliet demonstrated methods for normalising noisy tweets and categorising them based on polarity. The authors of this research used the Twitter search API to collect 2 million tweets from September 2009 to June 2010. They compiled a list of tweets about the mobile operation. They used a mixed model technique to generate sentimental terms, calculating the F-score of each word and selecting the words with an F-score more than 10% as raw words. They proposed a framework for gaining knowledge of the lexicon that can be taken from the gathered tweets so that we may express the terms in their most basic form as a future project. The GA studies begin with a vast amount of syntactic, semantic, and discourse level features that can be extracted. The fitness function determines the subjectivity classifier's accuracy based on the feature set selected by natural selection after each generation of crossover and mutation. For many domains, the ensemble technique, which combines the results of numerous basic classification models to generate an integrated output, has proven to be beneficial. In order to get better results, a new architecture based on coupling classification algorithms with an arcing classifier and suited to the sentiment mining problem is defined.

III. Proposed Methodology

1. Dataset

The reviews7 corpus was used for this project, and it contains 1000000 positive and 1000000 negative example movie reviews, each in an unprocessed HTML file. In addition, a custom dictionary is employed, which comprises all of the jargons as well as their definitions, and this, too, is used in conjunction with the training dataset. The benchmark dataset for sentiment analysis and domain adaptation is this dataset. We used Flickr's API service to search for images using each of the emotional categories - Fear, Happiness, Love, Sadness, and Violence - as search query parameters (along with the licence flag set to 'creative commons') to collect image metadata (such as server and farm ID numbers on which the image is stored) after sorting the result set by interest to ensure that images fitting the emotional categories are retrieved first. We gathered a total of 9854 photographs for this project, with 1900 images in each category. We divided the data in each category so that 75% of the data (8850) was used for training and 25% of the data (1000) was used for testing.

2. Data Pre-Processing

The most important stage in any Machine Learning model is data pre-processing. The model's performance is heavily influenced by how effectively the raw data is processed.

Text Pre-processing.

Text Processing is also the first step of Natural Language Processing. The following are the numerous pre-processing steps:

1. Lower Casing
2. Tokenization
3. Punctuation Mark Removal
4. Stop Word Removal
5. Stemming
6. Lemmatization

Text Pre-processing Using Lower Casing

We find terms in both lower and upper case when we have a text input, such as a paragraph. The computer, on the other hand, treats the identical words typed in different cases as distinct entities. For example, the computer treats 'girl' and 'girl' as two distinct words, despite the fact that they have the same meaning.

Tokenization

Tokenization is the next text pre-processing step. Tokenization is the process of dividing a paragraph into smaller parts like sentences or words. Each unit is then treated as a separate token. Tokenization's basic premise is to try to decipher the meaning of a text by studying the smaller units or tokens that make up a paragraph.

We'll utilize the NLTK library for this. The Natural Language Toolkit (NLTK) is a Python package for text pre-processing.

Punctuation Mark Removal

This brings us to the next step. We must now remove the punctuation marks from our list of words. Let us first display our original list of words.

Stop Word Removal

Stop words are a group of words that appear frequently in any language but add little meaning to sentences. These are frequent words that can be found in any language's grammar. Stop words are unique to each language. Stop words in English include "the," "he," "he," "his," "her," "herself," and so on. We can easily eliminate these stop words from our text data because they contribute little value to the overall meaning of the sentence. This aids in the decrease of dimensionality by removing redundant data.

Stemming

Stemming is the process of reducing a word to its root or stem word, as the name suggests. Only the root form or lemma remains after the word affixes have been eliminated. The words "connecting," "connect," "connection," and "connects," for example, are all reduced to the basic form "connect." "Studying," "studies," and "study" are all shortened to "studi." The word list generated by stemming does not necessarily include words that are part of the English lexicon. Terms like "scientif", "studi", and "everi" are not valid words in our example because they make no sense for the algorithmic context.

Lemmatization

We saw how we can use stemming to reduce words to their root words. Stemming, on the other hand, does not always yield terms that are part of the language's vocabulary. It frequently leads to terms that are meaningless to the consumers. To get around this limitation, we'll employ the concept of lemmatization. The words in our list have been lemmatized, as we can see. Every single word has been transformed into a meaningful parent word. Another important distinction between stemming and lemmatization is that lemmatization allows us to pass a POS parameter. By mentioning the parts of speech, we can establish the context in which we want to lemmatize our words (POS). The default is 'noun' if nothing is specified.

3.1 Features

Unigram

This feature is undergoing some pre-processing, such as stemming and stop word removal. As a result, we're looking at different types of unigram features, such as unigrams with stop words, unigrams without stop words, and unigrams without stemming with stop words. We don't include non-stemmed non-stop words in this group since non-stemming causes high dimensional features because it doesn't reduce words to their basic form.

Bigram of words:

We consider two types of Bigrams, similar to unigrams: Bigrams without stemming and stop words, and Bigrams with stemming and stop words. Bigram features have a high dimensionality. Features that appear more than three times in our dataset are taken into account.

Document Indexing

Indexing is the process of creating a feature vector or other representation of a document in the IR community. In these methods, a lexicon of words for the representations is defined, which includes all conceivable words that could be relevant to categorization. The value of each dimension (also known as an attribute) is allocated to a textual instance (say a document or a sentence) based on whether the term corresponding to that dimension occurs in the textual instance. The "bag of words" technique is what it's called. When the word is present, one essential concern is what values to employ.

The most frequent method is to weight each present word according to its frequency in the document and maybe in the entire training corpus. A binary weighting function is employed in most sentiment classification projects. It has been proven that assigning 1 if the term is present and 0 otherwise is the most effective method.

Information gain (IG)

In the world of machine learning, information gains are the most widely utilized feature selection strategy. It analyses the existence or absence of a feature in a document to determine its usefulness for predicting review sentiment.

Chi-square

The chi-square statistic determines how far expected and observed counts differ.

The frequencies W, X, Y, and Z indicate the existence or absence of a feature in the sample. W is the number of samples in which both features f and c appeared at the same time. We may also figure out what each symbol means by using TABLE 1. N=W+X+Y+Z. The feature is f, and the class is c.

Classification

In order to identify the ideal feature length, we are using Support Vector Machine to develop a classification model based on features of various lengths retrieved from these sorted lists. SVM can handle high-dimensional data in either a linear or non-linear way. SVM requires a long time to develop a classification model, yet it works effectively for two types of situations.

Image Analysis

With FER, the given image is fine-tuned. The confusion matrix shows that each of the good (Love, Happiness) and negative (Violence, Fear, Sadness) emotions has more confusion. We use sentiment analysis on our dataset to better comprehend our findings. The network learns to discriminate between the general positive and negative categories, but has problems categorizing within each of the positive and negative moods, resulting in a testing accuracy of 67.8%. We believe this is due to the network's learning of colours; positive sentiments are frequently associated with bright images, whereas negative sentiments are associated with dark images.

Cross Validation

Cross-validation, also known as rotation estimation, is a method for determining how well the results of a statistical analysis will generalize to a different set of data. It's most commonly employed in situations when the goal is prediction and the user wants to know how well a predictive model will perform in practice. The method of 10-fold cross validation is widely employed. The folds are chosen in stratified K-fold cross-validation so that the mean response value is about equal across all folds.

Criteria for Evaluation

Classification is the most important metric for evaluating classifier performance. Accuracy is the percentage of correctly categorized test samples. The capacity of a classifier to properly anticipate the label of new or previously unknown data is referred to as its accuracy (i.e., tuples without class label information). Similarly, a predictor's accuracy relates to how well it can anticipate the value of a predicted attribute for new or previously unseen data.

Applications:

Product Aspect Ranking, contextual analysis, ensemble feature selection and unigram method used together can provide very accurate results of sentimental analysis, as these methods together almost nullify all the drawbacks that are available in obtaining the classified output. Some applications where all these methods are used together are

Contextual Sentimental Analyser and Visualizer

The textual data is classified (single and multiple) and sentiment analysis is performed, with the textual data being scraped from microblogging sites or provided by the user. Furthermore, with the help of FER, photographs with clear faces can be analysed and emotions portrayed. The results are represented using visualization tools, which may be used to gain important insights that can then be leveraged to reduce customer churn and attrition.

Audio Parser

Voice input can be collected in real time or via pre-recorded recordings. After transcription, the audio and transcribed textual data is recorded, and sentimental analysis is performed on it. This paradigm can be applied to telemarketing and contact centres, for example.

System Architecture

We designed the system to improve sentimental analysis as well as Optical Character Recognition for forms and tables, making it more efficient and faster than existing systems. The suggested system also has a higher overall accuracy. The process in the proposed system is as follows: The system is fed input data that is either received from the user or scraped from the internet from microblogging sites. Instead of utilizing the pre-trained model TextBlob for sentimental analysis, a machine learning model is trained from scratch, and a python dictionary including all jargon is constructed, and the same dictionary is used for the dataset for handling jargon. To further analyse the context of the produced sentence and so obtain a better result, the Unigram approach and ensemble feature selection are utilized. In addition to the approaches outlined above, product aspect ranking is applied, and this hybrid methodology ensures the highest level of accuracy for text-based sentiment analysis.

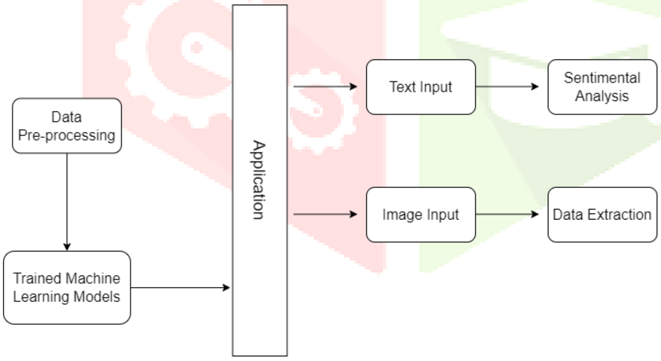

Fig.1 Proposed System Architecture

Experimental Results

The experiments are carried out on the data from the review. We add a pre-processing step to the dataset before doing sentiment classification. After pre-processing the samples in the corpus, the features are preferred.

Stemming is a type of pre-processing that is used to reduce inflected versions of words to their root form. The Porter Stemmer module, which employs the Porter stemming algorithm9, accomplishes this. We will utilize the selection criteria Mutual information, Information gain, chi-square, and TF-idf for feature selection. Each feature is given a score based on the two classifications, which are then ordered in decreasing order of their score. The SVM classifier is used to classify features of various lengths. The classifier is evaluated on feature sets ranging in size from 100 to 3000 of features (bag-of-words). We add a pre-processing step to the dataset before doing sentiment classification. After pre-processing the samples in the corpus, the features are preferred. Stemming is a type of pre-processing that is used to reduce inflected versions of words to their root form. The Porter Stemmer module, which employs the Porter stemming algorithm9, accomplishes this. We will utilize the selection criteria Mutual information, Information gain, chi-square, and TF-idf for feature selection. Each feature is given a score based on the two classifications, which are then ordered in decreasing order of their score. The SVM classifier is used to classify features of various lengths. The classifier is evaluated on feature sets ranging in size from 100 to 3000 of features (bag-of-words). In Tables 2 and 3, we show the accuracy of unigram features in both positive and negative classes. In the positive class, it was discovered that unigrams with stop words and unigrams without stop words have accuracy of 82.9 percent and 83 percent, respectively, with information gain. We also acquire good accuracy with knowledge gain in the negative class. The accuracy of a negative class unigram with stemming and no stop word is 83.1 percent. In both classes, Tables 4 and 5 show the accuracy of bigram with various feature selection criteria. When compared to bigram, the performance of unigram is superior. In both classes, bigram without stemming with a stop word gives superior accuracy. For both good and negative reviews, we indicate the accuracy (percentage) observed. We get the same accuracy for positive and negative reviews (53.4 percent and 53.5 percent, respectively) for all feature selection criteria. The outcomes for function words and Word-based measures were 64.4 percent and 67.8%, respectively. We observed that the unigram of bag of words is the best feature for extracting sentiment from reviews when we compared the performance of four features: unigram, bigram, function words, and POSTags of words. Previous research evaluated the results of several datasets and found that the ensemble feature set produced better results. The specified data set is being used to evaluate the performance of base and hybrid classifiers. The accuracy of the classification was tested using a 10-fold cross validation procedure. In the suggested method, the base classifiers NB and Genetic Algorithm are built separately to provide excellent generalization performance. Second, the NB, GA ensemble is created. The final output in the ensemble technique is determined as follows: the output of the base classifier is given a weight (0–1 scale) based on the generalization performance. The suggested hybrid NB-GA model improves classification accuracy substantially more than the base classifiers, and the results are statistically significant. In terms of classification accuracy, the suggested hybrid NB-GA algorithm outperforms individual approaches for film review data.

Conclusion

Our sentimental analysis system can collect omnichannel inputs and effectively classify them. To get the result, we employed machine learning algorithms like Naive Bayes Classifier and Support Vector Machine. Multiple approaches, such as ensemble feature selection, unigram method, and product aspect ranking, are included to identify the sentiment of the provided input, and because the algorithm consists of these methods, the accuracy of this system is superior than the previous ones. Furthermore, the output is presented in a variety of ways that may be used to derive important insights, making it easier for the user to manage the processed data and improving overall usability.

References

[1] J. C. Bezdek and R. J. Hathaway.: Convergence of alternating optimization. in Journal of Neural, Parallel & Scientific Computations, vol. 11, pp. 351-368. USA. 2019.

[2] C. C. Chang and C. J. Lin.: Libsvm: A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/simcjlin/libsvm/, 2018.

[3] G. Carenini, R. T. Ng, and E. Zwart.: Multi-document Summarization of Evaluative Text. in Proc. of ACL, pp. 3-7. Sydney, Australia. 2019.

[4] China Unicom 100 Customers iPhone User Feedback Report, 2016.

[5] ComScore Reports http://www.comscore.com/Press Events/Press Releases, 2014.

[6] X. Ding, B. Liu, and P. S. Yu.: A Holistic Lexicon-based Approach to Opinion Mining. in Proc. of WSDM, pp. 231-240. USA. 2018.

[7] G. Erkan and D. R. Radev.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. in Journal of Artificial Intelligence Research, vol. 22, pp. 457-479. 2014.

[8] O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates.: Unsupervised Named-entity Extraction from the Web: An Experimental Study. in Journal of Artificial Intelligence, vol. 165, pp. 91-134. 2015.

[9] A. Ghose and P. G. Ipeirotis.: Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Review Characteristics. in IEEE Trans. on Knowledge and Data Engineering, vol. 23, pp. 1498-1512. 2020.

[10] V. Gupta and G. S. Lehal.: A Survey of Text Summarization Extractive Techniques. in Journal of Emerging Technologies in Web Intelligence, vol. 2, pp. 258-268. 2019.

[11] W. Jin and H. H. Ho.: A novel lexicalized HMM-based learning framework for web opinion mining. in Proc. of ICML, pp. 465-472. Montreal, Quebec, Canada, 2019.

[12] M. Hu and B. Liu.: Mining and Summarizing Customer Reviews. in Proc. of SIGKDD, pp. 168-177. Seattle, WA, USA, 2018.

[13] K. Jarvelin and J. Kekalainen.: Cumulated Gain-based Evaluation of IR Techniques. in ACM Transactions on Information Systems, vol. 20, pp. 422-446. 2014.

[14] T. L. Wong and W. Lam.: Hot Item Mining and Summarization from Multiple Auction Websites. in Proc. of ICDM, pp. 797-800, Washington, USA. 2015.

[15] Y. Wu, Q. Zhang, X. Huang, and L. Wu.: Phrase Dependency Parsing for Opinion Mining. in Proc. of ACL, pp. 1533-1541, Singapore. 2019.

[16] J. Yu, Z.-J. Zha, M. Wang, and T. S. Chua.: Aspect Ranking: Identifying Important Product Aspects from Online Consumer Reviews. in Proc. of ACL, pp. 1496-1505, Portland, USA. 2014.

[17] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W. Y. Ma.: Improving Web Search Results Using Affinity Graph. in Proc. of SIGIR, pp. 504-511, Salvador, Brazil. 2015.

[18] Z. Zhang and B. Varadarajan.: Utility Scoring of Product Reviews. in Proc. Of CIKM, pp. 51-57. Arlington, USA. 2018