IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

NOVEL MACHINE LEARNING ALGORITHM FOR MULTICLASS SENTIMENT ANALYSIS

[1]Shubhrata Kanungo, [2]Suresh Jain

[1]Department of Computer Science & Engineering, Medi-Caps University Indore

[2]Department of Computer Science & Engineering, Medi-Caps University Indore

Abstract

Since the explosion of opinion-based web content, sentiment analysis has drawn the researchers ' interest in recent years as an extension of natural language processing. It area has brought a lot of growth which has encouraged the achievement of optimum classification of text results. We provide a collection of Advanced pattern-based applications to identify tweets along with other tools. In this article, together with their hybrid variations, we worked with the commonly used standard classifiers and machine learning to refine specific parameters so as to obtain the best possible classification accuracy. We performed our studies and provided relevant results and observations on the named film review corpus. The importance of opinion prediction via Facebook is investigated in this paper using machine learning approaches. Twitter-specific social network structures, like retweets, are also taken into account. Additionally, we are concentrating on locating both short and long phrases that are pertinent to the situation so that we can comprehend its effects. We used supervised machine learning methods such supporting vector machines (SVM), Naive Bayes, maximum entropy, and artificial neural networks to categorise data from Twitter using unigram, bigram, and unigram. The Bigram (hybrid) Extraction Model's Function.

Keywords: Twitter analytics, sentiment analysis, and support vector machines.

Introduction

In the simplest situation, the study of emotions leads to the function of binary sorting, discriminating between "strong" and "bad" mindset towards the topic under discussion. Multi-class classification arises when we include a "neutral" attitude, and regression issues are resolved when we express the mood in an ordered scale. It is possible to analyze not only the feeling itself but the author's emotions too. While the mood is generally graded as good, negative or neutral, feelings can be described on a wider scale. Paul_Ekman describes 6 basic emotions: joyful, pleased, delicate, scared, angry and sad. Plutchik, describes eight basic emotions as positive and negative in pairs: happiness versus sadness; rage versus fear; trust versus disgust; and excitement versus expectation. Many scholars also consider including audial and visual details for multimodal material sentiment analysis (Poria et_al), Consequently, sentiment analysis is not limited to text mining. Sentiment can be assessed at the textual, sentenceal, or dimensional levels. The fundamental structure of an emotional research is study at the document level. Here, the entire text serves to represent the author's viewpoint towards the subject under discussion. However, since a single document may include many viewpoints regarding the same thing, a study of the sentiment at the phrase level offers a clearer and more accurate picture. When an identifiable "atomic" object is referenced, sentiment analysis at the document or phrase level is successful. When evaluating product reviews for items with well-defined attributes that are being analysed by the reviewers, such as electronics, mobile phones, vehicles, or cameras, it is usual practise to evaluate emotions based on dimension.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Expressions of emotion are connected to several parts of the subject under discussion in this article, making it unnecessary to categorise the entire review as either good or negative for the company. Comparative sentiment analysis is an intriguing form of sentiment analysis where the reviewer compares the product to another "referential" product rather than expressing his feelings about a particular product (Feldman 2013). We must go through a number of procedures, including document collecting, document preparation, and feeling classification, in order to perform sentiment analysis automatically. Fig. 1 depicts a general approach for assessing emotions.

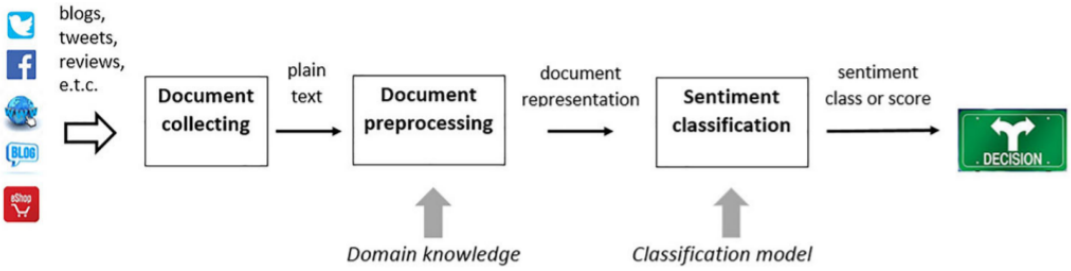

Figure 1: General process approach for assessing sentiment.

For various types of sentiment analysis (document-based, sentence-based, and aspect-based), various types of preprocessing and textual data encoding are required. a sufficient number of words for sentiment analysis at the document level. The terms used in the text are here displayed as a list. While avoiding the word's placement in the text, this representation acknowledges its frequency in the notion (a document is described in the database as a matrix of values extracted from the words' frequencies). After the bag-of-words composition, pre-processing operations are carried out, including word segmentation (to define distinct terms), lemmatization (to remove inflectional endings), stemming (to convert the words into their basic form), and stop-word elimination (to remove frequent words like conjunctions or prepositions not relevant to the document's content). We must first select the subjective sentences, or sentences that reflect the author's point of view, in order to analyse sentence-level emotions. Part-of-speech tagging, which labels word categories, can be applied in this situation to identify, for instance, adverbs closely connected to expressing opinions. Also, this step can be supported by various sentiment lexicons. Instead, similar to text-level analysis, sentences can be represented in a common way or characterised using words from meaning lexicons (this representation is also applicable at the topic level). Utilising knowledge extraction methodologies is crucial for an interpretation of feelings at the aspect stage. When the names of the items, suppliers, or product attributes must be included in the document, an explanation can be the identity of the designated person. The analysed data need not be sufficient in sentiment analysis, much like in a number of other text mining applications, and domain awareness is also crucial. When performing sentiment analysis, the subject matter data frequently includes a sentiment lexicon, which contains terms that describe various polarities of opinion. Terms are categorised into conditional (positive or negative) classes by a different group of lexicons; examples of such lexicons include LIWC or GI. A text analysis programme called LIWC comprises about 1000 words that, among other things, reflect good or negative emotions. (Tausczik_Pennebaker). An early lexicon with around 11000 words organised into 183 groups is called GI. This includes approximately 2000 terms that have been deemed optimistic and about 2300 words that have been classified negative (Stone_1966). Terms of opinion that were associated with the sentiment scores were included in the second batch of sentiment lexicons. SentiWordNet, a WordNet expansion in which around 147 000 WordNet synsets are connected with three number scores: Obj_(s), Pos_(s), and Neg_(s), denoting how factual, positive, and negative the words contained in the synset are (Esuli_Sebastiani), is a typical example

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

of this type of emotion lexicon. The second example of an emotion lexicon that links ratings with words is SenticNet. SenticNet is a freely accessible tool for concept-level study of perspectives and emotions; rather than offering vocabulary or single words, SenticNet delivers sentics (affective information) that are synonymous with 30,000 popular keywords. SenticNet, in contrast to many other emotion lexicons, incorporates graph mining and dimensionality reduction techniques to be immediately based on affective common sense awareness (Cambria_Hussain). For tasks related to sentiment analysis, domain knowledge is used in addition to feeling lexicons. Hand-crafted abstraction or grammar rules in NLP techniques are additional types of information employed in the field. (Jurafsky_Martin). The data (texts to be analysed) or domain experts may both provide the knowledge needed for automatic sentiment analysis. Various data mining (text mining) and machine learning approaches may be applied in the latter case. Having an adequate supply of labelled (annotated) training instances is essential in order to avoid the fundamental disadvantage of categorization using machine learning algorithms. It can be difficult to guarantee this for emotion classification because, similar to other text mining tasks, we have a lot less documents (examples) than terms (attributes) to work with. The most prevalent machine learning techniques are naive bayesian classifiers (NB), supporting vector machines (SVM), and pointwise mutual entropy (PMI), along with decision laws, k-nn techniques, and decision trees, according to a number of literature reviews. The NB classifier uses the independence assumption of input attributes ti (terms in document) given the class cj (sentiment direction) to determine whether a document represented by terms t1, t2,...,tn belongs to class cj based on the product of probabilities P (ti|cj) taking into account the contribution of each term ti independently on the others.

$$P(c_j|t_1, t_2, \ldots, t_n) = \frac{P(c_j)\prod_{i=1}^{n} P(t_i|c_j)}{P(t_1, t_2, \ldots, t_n)} \quad \ldots(1)$$

$$PMI(c_j, t_i) = \log \frac{P(c_j, t_i)}{P(c_j)P(t_i)} \quad \ldots(2)$$

We focus on rule-based and case-based solutions in this study because we believe that these interpretable models can provide more benefits than simple blackbox models.

The biggest obstacles to feeling analysis are as follows:

- Finding spam and phoney reviews: The internet has both legitimate and junk content. For the Sentiment to be correctly detected during transmission, such spam content should be eliminated. Finding duplicates, spotting outliers, and taking into account the reputation of reviewers can all help with this.
- Filtering classification while the definition or way of thinking that is most widely accepted is chosen is constrained. In order to improve the classification of feelings, this restriction needs to be reduced. The risk of a filter bubble produces unimportant groupings of opinions and a misrepresentation of mood.
- Asymmetry in the accessibility of opinion mining software: Because it is so expensive, only extremely large organisations and governments can currently afford it. That goes above and beyond what a regular person would anticipate. This should be available to everyone so that everyone can benefit from it.
- Integration of speech with implicit and behavioural evidence: To conduct a successful research of emotion, contextual information should be coupled with verbal expressions of emotion. The real behaviour of phrases with emotion is determined by the tacit data.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

- Domain-independence: The primary challenge for opinion mining and sentiment analysis is the availability of sentiment phrases in various domains. In one area, a particular set of features will perform very well, but very poorly in another.
- Natural language overhead processing: This includes overhead language features including ambiguity, co-reference, implicitness, inference, etc.

Related work

It is possible to see the categorization of sentences into positive, negative, and neutral categories as a natural language processing function. It has numerous levels of granularity, ranging from that of the document to that of the sentence [4]. Micro blogging platforms like twitter and facebook are populated with responses and views in real time. Numerous studies have been conducted in the past on sentiment analysis, mostly in the fields of product reviews, movie reviews, and political exit polls. Krishnamoorthy. The use of performance metrics for predicting the sentiment of financial texts is examined in this paper. This proposed a hierarchical classifier that used rules for associations as their definition. The efficiency of the classifier has been demonstrated by in-depth statistical analysis on a financial comparison dataset. Interesting results emerged from a study of how performance indicators affect the evaluation of emotions. Various data sets were discovered with varied degrees of lagging indicators, leading indicators, and emotion words' effects on the assessment of feelings.

Renault, T. (2019) Fnd that the addition of bigrams and emojis significantly improves performance in the classification of feelings. The performance of classification is not improved by time-consuming, more complex neural networks or other machine learning techniques like random forests. As a result of the size of the dataset and the preprocessing method, we also provide empirical support for the relationship between investor sentiment and stock returns. Despite the significant link between stock performance and investor sentiment. We do not observe any correlation between investor mood as expressed in social media posts and the regular return of high capitalization stocks.
Bhoyar, K_K_Bogawar, P_S., (2018) The data is sorted into two groups largely using Support Vector Machine (SVM). When utilising SVM to solve multi-category issues, researchers employed two different approaches. One approach focuses on the resolution of numerous SVM binary classifiers, whilst the other approach relies on the resolution of a single optimisation issue. In this article, the first approach was employed, and a warped binary tree-based Efficient Multiclass Support Vector Machine (ESVM) algorithm was suggested. There is no additional labour required to construct the warped binary tree as opposed to using the binary tree approach. Using the example data sets, the algorithm is evaluated, and the tests are contrasted with the two SVM multiclass approaches. The outcomes of the ESVM are compared against five methodologies for handling multiple binary SVM classifiers and four methods for handling a single optimisation problem. The comparative tests show that the ESVM performs better in terms of accuracy than other modern algorithms.

Tsakalidis , Baltas, A., Kanavos, A_ A. K. (2017) Due to the type, variety, and volume of the data, evaluating opinions based on Twitter data is a challenging problem. As part of this project, we are creating a machine using Apache Spark, an open-source big data processing platform. The sentiment analysis platform is built on machine learning approaches and makes use of Apache Spark's MLlib machine learning module, along with NLP techniques. We use a number of pre-processing techniques in the sentiment analysis to improve findings as a result of the complexity of big data. For binary and ternary classification, ranking algorithms are used. We examine the effects of sample size and input variables on test accuracy. Last but not least, the suggested system has been

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

trained and evaluated using real data that was scraped from Twitter, and the results are compared with those of actual users.

Language models and word vectorization

Nearly all NLP implementations depend heavily on word representations. Based on a thorough syntactic and semantic analysis of the given text documents, the phrases utilised as inputs to the NLP systems are frequently interpreted as indices or vectors and mapped in a high-dimensional vector space. The lexical interpretation features are represented by these real-valued term vectors. According to the order of their frequency, the terms that exist in our dataset were categorised. The most prevalent term in the corpus is represented by index 1, for example. Instead of using the entire corpus when parsing the phrase and indexing the terms, To train all of our networks, we picked the top 5000 terms. Over the years, a large number of neural network language models (NNLM) using distributed feature representation have been suggested [6–8]. These models successfully address the high-dimensionality problems that most statistical models, including interpolated n-gram models, have [6]. We coupled our networks' n-gram (unigram and bigram) models and noted variations in the accuracy level that transpired. A contiguous group of n terms is represented by an N-gram. Instead of grouping all the terms in the corpus, this may also encourage the most popular n-grams. Given the occurrences of the first (n 1) terms, it calculates the probability of the subsequent terms.

System, Scoring Model, and Sentiment Lexicon

Studies on sentiment analysis have shown that certain classifiers can be taught by supervised algorithms to predict unlabeled data by using labelled data. Despite the disadvantage of a dictionary-based technique being unable to handle domain and context-specific orientations, the effectiveness of this sort of sentiment analysis can be increased by creating domain-specific lexicons beforehand. Regarding dictionary origins, earlier studies on sentiment analysis used term-frequency-focused dictionaries as well as hand-picked dictionaries from well-known dictionaries like Harvard IV-41 and Loughran -McDonald2 [10], dictionaries built into websites as built-in sentiment tags, and dictionaries created based on a word's measured sentiment scores and the likelihood the stock price would rise or fall. Due to the fact that the characteristics of the target domain can affect how well sentiment is classified, few research have employed dictionaries with an emphasis on domain-specific lexicons. As a result, the hand-selected, based technique was employed in the present work to create domain-specific lexicons. examining feelings, the ultimate orientation of a sentence can be altered by positive and negative terms, which can also alter the overall orientation of a text. By determining the situation of the financial markets using grammatical norms, Das increased the effectiveness of a sentiment analysis. To improve sentiment analysis performance, a number of terms were included. Multi-word phrases were therefore employed to accurately acquire news articles, most of which were related to housing prices, in order to reflect the lexicons in the language of the present study. Textual, mathematical, and dictionary-based technologies are all used in natural language processing. The dictionary-based approach entails categorising the most important terms with a lexicon. It is also possible to determine whether a sentence is favourable or bad by looking at emotional research. Regarding the approach to figuring out the text's meaning orientation, which is explained by fusing words from the same emotion lexicon to find polarity words for each sentence. It was constructed using textual features and a sentiment-scoring reporting model that implemented the process of estimating stock price.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Table 1 Compares the literature on financial sentiment analysis

Literature_	The type of text	Dataset_ Dictionary	Features_	Method_	Objective_
Krishnamoorthy, S. (2017)	financial news articles	LM dictionary	words-in-a-bag or n-grams	machine learning-based approaches	Financial news item sentiment analysis using performance indicators
Yang, H.-F., & Seng, J.-L. (2017)	News and Housing Prices	Harvard IV-41 and Loughran–McDonald	House prices and news sentiet	analyse the correlation	Examining the Relationship Between News and Housing Prices Using Sentiment Analysis
Singh, R. (2017).	IMDB movie reviews	http://www. imdb.com. data set	extended words, emoticons, and words with the proper emotions	J48, BFTree, OneR, Naive Bayes,	Machine learning classifiers to improve sentiment analysis
Harish, U. C., and N. M. Dhanya (2018)	Tweet Data	lingo dictionary	n gram, Length of Tweet, Length of a Tweet, Hasht\gs:	SVM, Naïve Bayes classier and Decision tree	Sentiment Analysis of Twitter Data on Demonetization
EA_Kolog, CS_Montero,and T_Toivonen (2017)	Social Influence Analysis in Text	academic orientation data set	story corpus	MNB and J48 and SMO	Text sentiment and social impact analysis using machine learning
Mehta, A., Parekh, Y., & Karamchandani, S. (2018)	binary sentiment classification	Word2Vec.	opinin that is in line with one's emotions.	RNN, LSTM	Techniques for Sentiment Analysis Using Deep Learning
Yu, G., Gu, X., Li, Y., R., and O. Habimana (2019)	Twitter Data	Transformer-based encoder representations (BERT)	word embeddings that are specific to emotions	models of common sense, sentiment-specific word embedding, models of cognition-based attention,and models of reinforcement learning	Deep learning techniques for sentiment analysis

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Dictionary, pre-processing, sifting, and filtering

Data on house costs were obtained after removing news reports. An expert has provided a collection of domain-specific lexicons for searching news stories concerning property values. Each news article's title and content were examined as part of the screening process. The report was regarded as pertinent to housing costs if the title or body text contained any industry-specific words. The specialist's list of domain-specific lexicons was used to collect the pertinent news articles, and natural language processing was used to structure the posts. Special characters like hyperlinks to websites, segmented sentences, and stop words were first eliminated before converting the articles to TXT files and UTF-8 encode format. During the preprocessing of the papers, the published articles are added to a record of the corpus. Previous studies have shown that the bag-of-words paradigm is an easy and efficient way to convey texts. [15, 30]. Hence, the translated news articles were also provided using this approach. After the above step each news article was translated to a lexicon frequency vector space, thereby quantifying the articles ' qualitative data. In order to improve the accuracy of the vocabulary utilised in this study, representative news articles were examined. Using the frequency-based technique, two indicators were chosen to order the news article in the corpus database by Berry and Kogan[2]. The application of the two predictor values allows the selection of the papers for relevant documents in descending order and scanning. According to Biber_etal.[3], establishing a high threshold value is beneficial for the analysis of numerous articles.

For the representative news articles, then, a top-30 question form was used to view. According to Nation_[29], who first developed a domain-specific dictionary, the present study has established a dictionary that combines two different lexicon types (domain-specific lexicons and emotion lexicons). The specific lexicons for each subject were also given to the expert. In the domain-specific lexicons, nouns constituted the majority of the words. The domain-specific lexicons were chosen with a focus on the characteristics of phrases connected to housing prices and included words relating to people, activities, period, places, and things. The sentiment lexicons are used to ascertain a sentence's sense orientation as opposed to domain-specific lexicons. Due to the greater level of detail they provide than single words, multiword phrases are helpful in detecting the meaning orientation of sentences. Thus, in this investigation, the lexicons of emotion were constructed using multiword phrases. The chosen domain-specific and emotion lexicons were then categorised, condensed, and revised. By feeling about news, categorising lexicons can improve the explanability of property prices. As a result, whereas the lexicons of emotion were only split into two groups (positive and negative), the gathered domain-specific lexicons were classified into a wide range of categories.

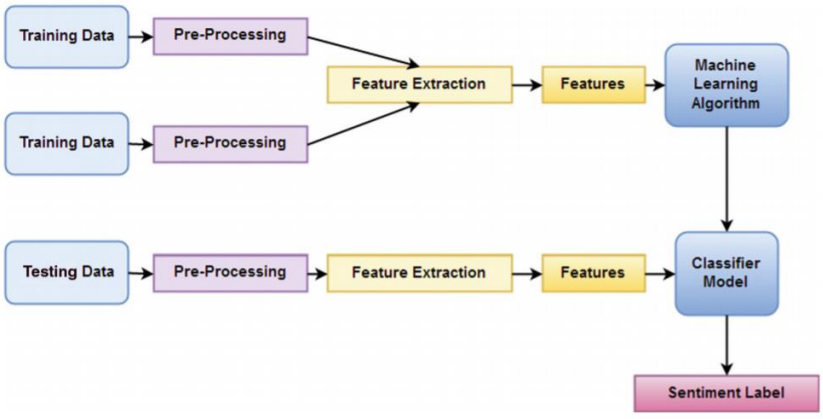

Figure 2: Divides the data into train & test, pre-process, feature selection, ML algorithm and finally label sentiment.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Typical Machine Learning Techniques

Among the current, widely used classification approaches are Bayesian classifiers like Naive Bayes[9], Maximum Entropy Classification[10], Technical Regression, and Vector Machine Assistance (SVM). These algorithms employ a standard bag-of-functions technique that categorises the text into a class or group c using a set of predefined "m" attributes found in document d.

Support Vector Machines (SVM), V. Vapnik introduced a supervised classifier in 1995[11] and shown that it was more reliable and effective in comparison to Support Vector Machines (SVM), which he had previously developed. Are simple tactics still more effective than machine learning methodology? In this article, The four machine learning algorithms that we take into consideration are the Naive Bayes algorithm (NB), the Maximum Entropy classifier (MaxEnt), the Vector Support Vector classifier (SVC), the Random Forest (RF) classifier, and the MultiLayer Perceptron (MLP) classifier. We optimise hyperparameters using the scikit-learn tool and a grid search. With the exception of using triple cross-validation and implementing a minimum word frequency of 0.0001 percent to exclude very uncommon terms and shorten computation time (GridSearchCV is very time-consuming), we focus on unigram and bigram text functions, including emojis and punctuation, on a structured dataset of 250,000 messages. We also provide the time required by each approach (for the largest number of hyperparameters) to complete the classification. We find that less complicated methods like Support Vector Machine or Maximum Entropy perform better in terms of classification accuracy than more complicated ones like Random Forest and Multilayer Perceptron. We believe that a straightforward classifier like Naive Bayes, a supporting vector machine, or maximum entropy, as in Antweiler and Frank (2004) and Sprenger et al. (2014), will frequently do the trick for analysing social media feelings given the costs of optimising the hyperparameters (time, complexity, lack of transparency, and cost of computing power).

Proposed algorithm

Advantages of Algorithm (Semi supervise Support vector machine)

1. The approach does not require any additional overhead to generate the binary tree, similar to S3VM-BDT.
2. The number of binary SVM classifiers as for S3VM-BDT[5] and S3VM-RH is reduced to N-1.
3. In the testing step, just one binary classifier is evaluated if the data sample belongs to Class 1. Otherwise, several classifiers are needed, like in the case of S3VM-BDT[5].
4. The decision function for each class is tested by S3VM-RH[9] in order to establish the data sets, and an object belongs to the class with the greatest decision function value when fewer decision functions are being evaluated in this process. Only the last level measures both judgement responsibilities.

The Sentiment Analysis Tool in this study was made using three categorization algorithms. We assessed the identification of binary and ternary on various datasets. The implications of the various characteristics of the vector that is used as input to the classifier will be the focus of our attention for ternary classification rather than binary classification, which focuses on how the sample size influences the tests. Naive Bayes, Logistic Regression, and Decision Trees are the three methods employed.

Basic Bayes Naive Bayes is a straightforward multiclass classification method that focuses on the Bayes theorem's interpretation. In that particular case of the question, which is specified as a function matrix, each feature's value is assumed to be independent of the values of all other features.This method has the benefit of only requiring a little amount of exposure to training data in

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

order to be learned. The conditional probability distribution of each specified class feature is first calculated using the Bayes' theorem in order to predict the class label for a particular instance.
Regression into Reason The dependent variable may take one of a preset number of values in a regression model called logistic regression. To ascertain how the instance class and the features gathered from the data interact, a logistic technique is used. It can be generally used to overcome problems with multiclass classification, while being typically used for binary classification.

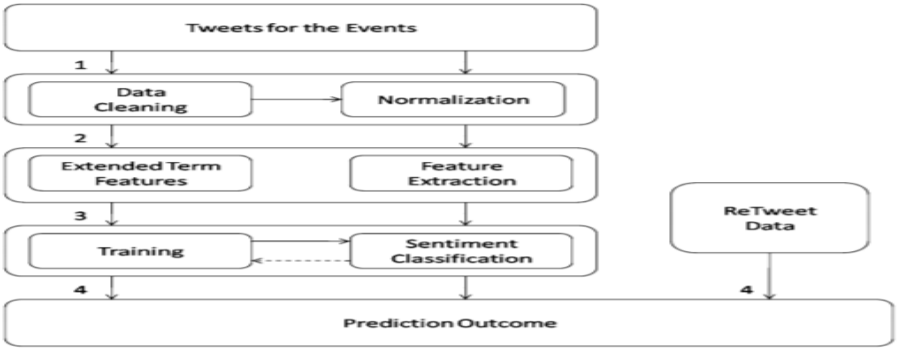

Figure 3: Proposed approach flowchart of work.

Suitable Trees A "decision tree" is a classification method that is based on a tree structure with leaves that indicate class labels and branches that reflect feature combinations that lead to the aforementioned classes. In essence, the feature space is divided into recursive binary partitions. To maximise the choice for the stated motion by maximising the benefit of knowledge, each step is carefully picked.
After that, each tweet is parsed and tokenized. Once special tags have taken the role of username and URL occurrences, the vector that represents each tweet then has the properties stated below:
– Unigrams, or word frequencies, are a common occurrence in tweets.
– Bigrams are two-word phrases that frequently appear in tweets.
– Trigrams are three-word phrases that are frequently found in tweets.
– The binary flag Username, which denotes if a user is named in a tweet.
– Hashtag, a binary flag that shows whether a hashtag is present in the tweet,
– URL, a binary flag that indicates whether a URL is present in the tweet.
– The following POS Tags are counted because they were applied to tokenized tweets using the Stanford NLT MaxEnt Tagger:

The quantity of adjectives, the number of verbs, the quantity of nouns, the quantity of adverbs, the quantity of interruptions, the number of interjections. Next, ratios are determined between the aforementioned numbers and the total number of tokens in each tweet.

Proposed algorithm:

Our objective is to gauge how a Twitter community as a whole feels about a certain subject. The first stage is to gauge how the neighbourhood feels about each individual tweeted message. We build a sentiment classifier using the training data and do sentiment analysis on Twitter. We use the Support Vector Machine (SVM) algorithm[12], and SVM performance[21–23] in particular. A labelled set of cases is necessary for the SVM method to build a model. Additionally, we obtained three Twitter datasets that were labelled, each of which had a different size, conversational focus, and labelling approach. We trained three connected sentiment models on the same testing set and evaluated their performance. The appropriate sense classification type is subsequently applied in the vast majority of our analyses. The initial dataset included 1.6 million tweets that Stanford University has classified as either good or negative [13]. The tweets are classified as either positive or negative

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

based on the presence of emoticons, which were later taken out of the training dataset. Although the labelling produced by this method isn't the greatest possible, it is a reasonable and cost-effective substitute for hand tweet labelling [24]. This dataset includes a variety of tweets that don't focus on any one issue. The tweets in this dataset are all-encompassing and unspecific in nature. General English messages were still present in the second sample, but the message names were obtained by manual coding. This dataset contains 37,951 neutral, 25,721 neutral, and 23,250 neutral handwritten tweets. Because they are a randomly chosen subset of our environmental tweets, the third dataset includes tweets that are largely domain-specific. 11,439 neutral, 2,850 neutral, and 5,569 neutral hand-labeled tweets from January to December 2014 are included in this dataset. To test the qualified emotion models, we choose randomly 20 percent of these tweets, while maintaining the marking distribution across the full dataset. 80% of the tweets were used to train the domain-specific sentiment model from the remaining domain data.

Positive and negative tweets are the only sources for emotion models. Positive, negative, and neutral are the three subcategories included in the classification. A tweet is categorised as positive (negative) if its distance from the SVM hyperplane is greater than the mean distance from the hyperplane of positive (negative, respectively) training examples. It is considered undesirable if it is too close to the hyperplane. Similar methods for applying the discrete SVM classifier to the environment of three groups have also been used in our earlier studies[24, 25]. Both common parsing techniques and Twitter-specific ones are effective for parsing tweets. Tokenization, stopping, the construction of unigrams and bigrams, the eradication of words that do not appear at least twice in the corpus, and the creation of word frequency (TF) attribute vectors are all parts of conventional processing before. While Twitter-specific preprocessing [8, 13, 24] reduces superfluous messages and converts usernames and hashtags. Using the corresponding positive and negative pre-processed messages, We developed three sentiment models (hand-labeled domain-specific, smiley-labeled general, and hand-labeled general) and evaluated the performance of each on the distinct test set mentioned above. The outcomes are shown in Table 1 based on the macro-averaged error rate[27] and the macro-averaged positive and negative class F-score[28]. Accurately classifying the good and negative tweets is something we are particularly interested in. Table 1 shows that the hand-labeled domain-specific sentiment model, which had the lowest error rate and greatest macro-averaged F-score on the test set, had the best performance.

Be aware that whereas the general model tagged with a smiling used 1.6 million tweets to train and the general model labelled with hands used 48,971 tweets to train, this model was only trained on 6,735 tweets. Therefore, even if there are fewer high-quality domain-specific tweets, the results show that they yield stronger sentiment models. The hand-labeled domain-specific sentiment model included with the entire hand-labeled domain-specific dataset is what we use for the remainder of our research. The following formula is used to determine how various communities feel about a certain subject. The tweets that its users post are first chosen for each community. Second, the number of retweets each tweet receives determines and weighs its sentiment. Third, the community as a whole totals the weighted positive and negative sentiment of each user's tweets. In the end, a community's predisposition towards a specific topic is calculated by multiplying the polarity of the combined weighted sensation by the proportion of tweets from that group that reflect emotion (subjectivity). The subjectivity and polarity measurements were extracted from [29]. The provided community sentiment computation pseudo-code

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Algorithm 1

Require: $\mathcal{C}$: community,
T_S : sentiment annotated tweets,
$\bar{D}_P$: avg. distance of positive training examples,
$\bar{D}_N$: avg. distance of negative training examples

function COMMUNITYSENTIMENT ($\mathcal{C}$, T_S):
$pos = 0$
$neg = 0$
$all = 0$
for *user* in $\mathcal{C}$.*users* **do**
userTweets = T_S.byUser(*user*)
for *tw* in *userTweets* **do**
if *tw.sentiment* > $\bar{D}_P$ **then**
pos += *tw.retweetCount*
else if *tw.sentiment* < $\bar{D}_N$ **then**
neg += *tw.retweetCount*
end if
all += *tw.retweetCount*
end for
end for
$polarity = \frac{pos-neg}{pos+neg}$
$subjectivity = \frac{pos+neg}{all}$
return *polarity* × *subjectivity*
end function

Conclusion

An excellent resource for assessing the objectivity and consistency of the public's thoughts is a social network like Twitter that permits public emotion. We evaluated many supervised classifiers for sentiment analysis over Twitter data utilising different feature extraction methods, such as unigram, bigram, and hybrid (unigram? bigram). By modifying the present extended functionality paradigm to eliminate the phrases that drive the case, we established a novel strategy. For assessing the effect factor generated by retweets on Twitter, we have proposed a methodology. The dataset's mood towards the current demonetization issue is evaluated. Though there are some opposing views, the analysis overwhelmingly portrays support for the change. The tests are compared using several classification machine learning algorithms[16]. Many researchers are interested in using machine learning approaches for sentiment analysis. As a result, numerous machine learning models have been proposed and have proven to deliver successful results on a variety of sentiment analysis tasks. The success of the offered solutions is a result of both their success in embedding verb model models and their capacity to learn automated function.

The backdrop of sentiment analysis, including its procedures, stages, and tasks for evaluating emotions, is covered in the first section of this essay. We also provide an overview of conventional methods for assessing emotions and their drawbacks. The most accurate model for our dataset is SVM. Our method's main flaw is that many are written in regional tongues. Our proposed algorithm achieves a fair degree of forecast accuracy for a participating party by equally considering the emotions of each Twitter user. It might be possible to give a certain famous user's opinions more weight than those of other Twitter users, which could improve the influence factor's ability to anticipate future events. The potential expectation for text-based sentiment analysis should be fruitful and quick given the exponential growth of Twitter data. In order to give high-performance, real-time computation results for predicting any event, hybrid classifiers must also be parallel.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

References

1. Krishnamoorthy, S. (.2017). Sentiment analysis of financial news articles using performance indicators. Knowledge and Information Systems, 56(2), 373–394. doi:10.1007/s10115-017-1134-1
2. Yang, H.-F., & Seng, J.-L. (2017). Using Sentiment Analysis to Explore the Association Between News and Housing Prices. Lecture Notes in Computer Science, 170–179. doi:10.1007/978-3-319-54430-4_17.
3. Singh, J., Singh, G., & Singh, R. (2017). Optimization of sentiment analysis using machine learning classifiers. Human-Centric Computing and Information Sciences, 7(1). doi:10.1186/s13673-017-0116-3
4. Dhanya, N. M., & Harish, U. C. (2018). Sentiment Analysis of Twitter Data on Demonetization Using Machine Learning Techniques. Lecture Notes in Computational Vision and Biomechanics, 227–237. doi:10.1007/978-3-319-71767-8_19
5. Berka, P. (2020). Sentiment analysis using rule-based and case-based reasoning. Journal of Intelligent Information Systems. doi:10.1007/s10844-019-00591-8.
6. Kolog, E. A., Montero, C. S., & Toivonen, T. (2018). Using Machine Learning for Sentiment and Social Influence Analysis in Text. Advances in Intelligent Systems and Computing, 453–463. doi:10.1007/978-3-319-73450-7_43.
7. Mehta, A., Parekh, Y., & Karamchandani, S. (2018). Performance Evaluation of Machine Learning and Deep Learning Techniques for Sentiment Analysis. Information Systems Design and Intelligent Applications, 463–471. doi:10.1007/978-981-10-7512-4_46 . Renault, T. (2019). Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. Digital Finance. doi:10.1007/s42521-019-00014-x
8. Renault, T. (2019). Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. Digital Finance. doi:10.1007/s42521-019-00014-x
9. Bogawar, P. S., & Bhoyar, K. K. (2018). An improved multiclass support vector machine classifier using reduced hyper-plane with skewed binary tree. Applied Intelligence. doi:10.1007/s10489-018-1218-y
10. Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016). A hybrid approach to sentiment analysis. 2016 IEEE Congress on Evolutionary Computation (CEC). doi:10.1109/cec.2016.7744425
11. Sorvisto, D., Cloutier, P., Magnusson, K., Al-Sarraj, T., Dyskin, K., & Berenstein, G. (2018). Live Twitter Sentiment Analysis. Lecture Notes in Social Networks, 29–41. doi:10.1007/978-3-319-95810-1_4
12. Baltas, A., Kanavos, A., & Tsakalidis, A. K. (2017). An Apache Spark Implementation for Sentiment Analysis on Twitter Data. Lecture Notes in Computer Science, 15–25. doi:10.1007/978-3-319-57045-7_2
13. Habimana, O., Li, Y., Li, R., Gu, X., & Yu, G. (2019). Sentiment analysis using deep learning approaches: an overview. Science China Information Sciences, 63(1). doi:10.1007/s11432-018-9941-6.
14. Sluban, B., Smailović, J., Battiston, S., & Mozetič, I. (2015). Sentiment leaning of influential communities in social networks. Computational Social Networks, 2(1). doi:10.1186/s40649-015-0016-5.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

15. Fersini, E. (2017). Sentiment Analysis in Social Networks. Sentiment Analysis in Social Networks, 91–111. doi:10.1016/b978-0-12-804412-4.00006-1.
16. Anjaria, M., & Guddeti, R. M. R. (2014). A novel sentiment analysis of social networks using supervised learning. Social Network Analysis and Mining, 4(1). doi:1