Computers, Materials & Continua
DOI: 10.32604/cmc.2023.032173
Article

Aspect Level Songs Rating Based Upon Reviews in English

Muhammad Aasim Qureshi[1], Muhammad Asif[2], Saira Anwar[3], Umar Shaukat[1], Atta-ur-Rahman[4], Muhammad Adnan Khan[5,*] and Amir Mosavi[6,7,8]

[1]Department of Computer Science, Bahria University Lahore, 54000, Pakistan
[2]Lahore Institute of Science and Technology Lahore, 54792, Pakistan
[3]Department of Multidisciplinary Engineering, Texas A&M University, College Station, 77843, USA
[4]Department of Computer Science, College of Computer Science and Information Technology (CCSIT), Imam Abdulrahman Bin Faisal University (IAU), P.O. Box 1982, Dammam 31441, Saudi Arabia
[5]Department of Software, Gachon University, Seongnam, 13120, Korea
[6]John von Neumann Faculty of Informatics, Obuda University, Budapest, 1034, Hungary
[7]Institute of Information Engineering, Automation and Mathematics, Slovak University of Technology in Bratislava, Bratislava, 81107, Slovakia
[8]Faculty of Civil Engineering, TU-Dresden, Dresden, 01062, Germany
*Corresponding Author: Muhammad Adnan Khan. Email: adnan@gachon.ac.kr
Received: 09 May 2022; Accepted: 12 June 2022

Abstract: With the advancements in internet facilities, people are more inclined towards the use of online services. The service providers shelve their items for e-users. These users post their feedbacks, reviews, ratings, etc. after the use of the item. The enormous increase in these reviews has raised the need for an automated system to analyze these reviews to rate these items. Sentiment Analysis (SA) is a technique that performs such decision analysis. This research targets the ranking and rating through sentiment analysis of these reviews, on different aspects. As a case study, Songs are opted to design and test the decision model. Different aspects of songs namely music, lyrics, song, voice and video are picked. For the reason, reviews of 20 songs are scraped from YouTube, pre-processed and formed a dataset. Different machine learning algorithms—Naïve Bayes (NB), Gradient Boost Tree, Logistic Regression LR, K-Nearest Neighbors (KNN) and Artificial Neural Network (ANN) are applied. ANN performed the best with 74.99% accuracy. Results are validated using K-Fold.

Keywords: Machine learning; natural language processing; songs reviews: sentiment analysis; songs rating; aspect level sentiment analysis; reviews analysis; text classification; music

1 Introduction

Since last years to interact between social users the internet has gained popularity and becomes and backbone of social media [1]. It has digitized the mechanical world [2]. Everyone has quick accessibility to portable devices that have a stable internet connection. People are using the internet

for business and social correspondence [3]. In past, there were some specialized companies to collect reviews and feedback, for decision making, regarding the product in hand through the market survey but it is an old-fashioned way to collect feedback [4]. People are becoming habitual in buying and selling products online. After online shopping, they post their experience (in the form of a review) by rating the product or commenting on it [5]. These reviews i.e., feedbacks are important for both users—buyer and seller, as well. These reviews, on one hand, help the users, i.e., consumers, to understand the quality traits of the product and on the other hand, help the organization, in making decision, to improve its quality standards according to the users' needs [6]. Due to the exponential increase of e-users, these reviews are increasing day by day. In order to analyze these reviews and establish a user-centric rating, a complete automated mechanism is required. Internet has played an important role in rapid popularity and growth of entertainment industry [7]. This paradigm has provided easy access to all kinds of media like movies, plays, songs and many more, to the people. YouTube is one of the most important sources to access this kind of content. Nowadays YouTube [8] is a popular platform to host such material. The quality of the content is generally judged [9] on the basis of likes and dislikes on the content [10]. The content credibility can easily be judged by simple formula given in Eq. (1).

$$CQ \leq 0\,?\,(\text{“good”}:\ \textit{Bad}\text{”}) \tag{1}$$

where,

$CQ = TL - TD$

The Content Quality (CQ) uses (TL) Total number of likes (TL) and (TD) the Total number of dislikes.

To measure the credibility of the substance, it is not a good quality metric because it provides limited insight into the content. A better way to check the credibility of content is analyze the comments/reviews on that content [11]. The quality of the content can easily be judged by reading the reviews manually if these reviews are small in number. But it is humanly not possible when these numbers are large in number. This enormous increase in reviews demands an self-sunning mechanism to analyze these reviews. Regarding this, SA plays an imperative role in analyzing human sentiments present in the text [12]. Sentiment analysis is a way to analyze the sentiment of the users into positive, negative or neutral, hidden in text [13,14]. There are three level of sentiment analysis—document level, sentence level and aspect level. In the document level, whole the document is taken as single entity and analyzed. Sometimes, results do not endorse the actual expressions by text [15–17]. To conceal this, at the sentence level, the document is broken into sentences and each sentence is taken as one entity and analyzed. But there exists a dilemma if different aspects of the product are deliberated in a single sentence have minor differences but they have reverse meaning. To overcome these types of issues, sentiments are analysed at the aspect level [18–20]. At aspect level sentiment analysis each aspect/feature of the product is taken as an entity and can be analyzed. The main focus is to analyze the review at the aspect level. Aspect based sentiment analysis has a wide range of applications in different fields, like song reviews, hotel reviews, movies reviews, songs reviews and much more [21–23]. As most of the prior work is done at the document and sentence level in the field of sentiment analysis. These provide limited insight into the feature of the product [24]. Limited research is witnessed at aspect level sentiment analysis, so there is a need to extend research on aspect level sentiment analysis to improve different business strategies and customers need.

The western music industry carries a good marketplace and it is rising gradually. Internet technology plays a vital role in its enrichment. Reviews/comments in the entire song help to understand

the quality of content provided in the song [25]. A lot of work is witnessed on sentiment analysis on document and sentence level, very limited research is witnessed at aspect level sentiment analysis. Research on aspect level sentiment analysis of songs at the initial stage [26], so the scope of enhancement exists in that area. Therefore, there is a need to explore different aspects of songs and carry out research on aspect level sentiment analysis. This paper contributes a benchmark dataset named Corpus for Aspect Level Sentiment Analysis of Songs (i.e., CALSAS). The entire corpus consists of 369,436 reviews that belongs to three classes (positive, neutral and negative). CALSAS belongs to five different aspects (Lyrics, Music, Song, Video and Voice) called sub-dataset e.g., sub-dataset—voice. The entire corpus is annotated using the technique presented in [27]. Partially, this research is conducted to perform aspect level sentiment analysis of reviews. As well as this study recommends model to rate and rank of any item (i.e., rate the songs based on their aspect level sentiments) whose reviews are available. This model is designed and test against the dataset CALSAS. The rest of the paper is organized into the five-section. In Section 2 state of the art literature review is discussed in detail. The methodology is discussed in Section 3 which is adapted to achieve the objectives of the study. In Section 4, Modelling and Experimentation is discussed in length. Finally, the study is concluded with the direction which will be made in future.

2 Literature Review

In previous studies, efforts have been made to perform sentiment analysis on document level and sentence level. At aspect level sentiment analysis limited research is witnessed.

In [28] sentiment analysis of mobile reviews was performed at aspect level. The dataset that was collected from Amazon on three products consists of 1,350 sentences. Their proposed model takes as input a sentence and identify the Noun phrases as an aspect. The proposed model achieved F-score of 0.80% and precision was 78.0% and recall was 77.0%. In [29] sentiment analysis was performed on aspect level on customers reviews using lexical resources by applying supervised machine learning techniques. The dataset consists of restaurant and laptops reviews. At aspect "extraction", the maximum F1 is at restaurant data was 80.19% and on laptop reviews, the F1 was 68.57. At aspect "polarity" NRC gained 80.16% results in terms of accuracy at the reviews of laptops and NRC value is 88.58% in terms of F-score on restaurant reviews. In [30] active learner utilizing Lexicalized Dependency Path (LDP) is proposed to provide additional flexibility to the developers over the extraction model. The model is implemented in real active learning and simulated. This study improves 4.5% over coarse classes. In [31] sentiment analysis was performed to review a product at aspect level using supervised classifier Naive Bayes. Dataset used from different domain food, services, price, ambience and miscellaneous and the total number reviews were 3714. Results were calculated using two different methods by applying Naïve Bayes. First, the results were calculated without using Chi-square and the accuracy was 95.87%, after that the results were calculated with Chi-square, where the value of Chi-Square was 0.2 and the accuracy that was achieved 92.86%. For feature selection, POS tagging and Chai-Square method were adopted.

In [32] the strategy was adopted to perform sentiment analysis on English tweets using Naïve Bayes. Dataset conations 6408 tweets, the model achieved 0.63% in terms of accuracy. In [33] Sentiment analysis was performed at aspect level on e-commerce dataset by using supervised machine learning techniques i.e., Naïve Bayes and Support Vector Machine. The dataset was collected from Amazon Web and features were selected using POS tagging. The model was evaluated using an F-score, where the value of F—score 95.2% by naïve bayes and support vector machine performed 84.1%. In [34] the sentiment analysis was performed on online tourist reviews at aspect level. The dataset consists of

2000 restaurant reviews and 4000 hotel reviews. Aspects identification is performed by using implicit, explicit and co-referential techniques. The proposed model correctly classified 82% aspects in the hotel dataset and 91% in the restaurant dataset. In [35] sentiment analysis was performed on amazon reviews at aspect level. A supervised machine learning classification algorithm has been used to rate the products. TF/IDF and POS tagging techniques were used to extract features. In [36] recommendation of behaviors with the help of aspect level sentiment analysis. The dataset that was used consists of 2590 hotel and cars reviews and 3700 mobile reviews (Galaxy S8). The proposed approach performs 84.00% and by using POS F-score 78.76% at cars dataset. In [37] aspect base analysis of sentiments was performed using machine learning techniques of student opinions. Dataset consists of 1728 records was extracted by using API from twitter and sentiment analysis were performed by applying Naïve Bayes and SVM. Features are extracted by using POS tagging. Seven different aspects were extracted (teaching, placement, facilities, sports, organizing events, fees and transport). Naïve Bayes outperforms the other classifier and the value of the F-score was 0.987. In [38] sentiment analysis was performed on reviews of smart government datasets. Dataset was presented by lexical resources that were gathered from the Government apps and consist of Government app-specific reviews. Dataset consists of 7,346 number of examples. Feature-based SVM perform 82.45% accuracy and Lexicon based SVM performed 88.41%. The proposed model helps to improve the performance of government smart applications and their services by analyzing the sentiments of mobile apps reviews. In [39] sentiment analysis was performed 3,057 reviews on different products available on Amazon web i.e., mobiles, laptops, tablets, cameras, video and televisions. Using machine leering classification, Naïve Bayes outperforms the other classifiers and accuracy was 98.17% whereas the accuracy of SVM was 93.54%.

In [40] customer feedback analysis was performed. Reviews were collected from Amazon web, which was 500 in numbers, from different products i.e., computers, mobiles, flash drives and electronics. The products were categorized by applying POS tagging. In [41] an aspect-based sentiment analysis was performed using word-based, syntax-based and grammar-based features. A restaurant dataset has used that consist of 350 examples. Total 2499 number of features were extracted. SVM performed 72.4% in terms of accuracy. In [42] sentiment analysis of online reviews was performed on the Yelp dataset at aspect level, (i) Laptop reviews and (ii) restaurant reviews. A total 4,934 number of reviews were collected. At entire dataset Feature + SVM performed 72.10% at restaurant reviews and 80.89% on laptop reviews. Their proposed model performs 68.34% on restaurant data and 70.90% on the laptop dataset. Most of the research in English text sentiment analysis is done at document-level sentiment analysis and sentence-level sentiment analysis. Limited research is witnessed at aspect level sentiment analysis. Some of the work on aspect level sentiment analysis is observed, but it is not on a large-scale dataset i.e., dataset of 1350 sentences used in [28], the dataset of 800 restaurant reviews used in [30], the dataset of 3714 reviews on different products were used in [31], the dataset of 6408 tweets was used in [32] and other are listed in the literature. So, there is a need to extend a benchmark dataset of English text to extend research on aspect level sentiment analysis.

3 Materials and Methods

In this section, the methodology of the paper which is adopted to rate the songs based on their aspects and to perform aspect level sentiment analysis of songs is discussed in detail.

3.1 Dataset Collection

For any analysis data is an important aspect. It is not possible to perform any analysis without data. To build a gold standard dataset and for data collection, 10 top rated English songs are selected available on Kwrob [27]. Reviews are collected from YouTube. In the Tab. 1 selected songs and their singers' names are demonstrated. After scraping these reviews, save them into CSV file format. The scraped reviews are in ten different files, carrying targeting aspects and a lot of noisy data as well. To make data ready for the analysis and to separate the selected aspects different pre-processing techniques are applied.

Table 1: Number of reviews scraped from songs

No	Singer	Country	Title of song	Views	Comments
1	Justin Bieber	Canada	Sorry	3,203,542,747	816,063
2	Katy Perry	USA	Roar	2,932,210,456	638,490
3	Ed Sheeran	UK	Shape of You	4,449,412,646	910,006
4	Taylor Swift	USA	Shake It Off	2,832,020,062	517,249
5	Shakira	Colombia	Chantaje	2,455,461,250	380,955
6	Rihanna	USA	Calvin Harris-This Is What You Came	2,275,349,039	287,379
7	Eminem	USA	Love The Way You Lie ft.	1,868,708,629	523,115
8	Natti Natasha	Dominican	Ozuna Criminal	1,889,274,387	259,644
9	Maroon 5	USA	Sugar	3,047,850,854	342,142
10	Enrique Iglesias	Spain	Bailando ft. Descemer Bueno, Gente De Zona	2,774,918,985	211,363

3.2 Preprocessing

Noisy/unprocessed data lead to unreliable results because the results of any analysis are directly affected by the quality of the data [43]. To avoid erratic results, different preprocessing techniques are applied to get consistent results. As discussed above the scraped reviews contains reviews of targeted aspects as well as other reviews with a lot of noisy data. To preprocess and to get the reviews that contains targeted Aspects, this study adopted various preprocessing techniques filtration of targeted aspects, convert the uppercase into lowercase, remove the emoji's and the reduction string size.

3.2.1 Aspect Filtration

The collected reviews contain reviews on different aspects which covers the different features of the songs and reviews in other languages as well. For this research, five Aspects (music, lyrics, song, voice and video) are chosen to perform aspect level sentiment on song reviews. Though there exist techniques for aspect filtration like N-gram, TF/IDF, etc. but for this study, a survey was conducted from 20 people (10 male, 10 female). They were given options against songs in general to rank each option from 1 to 10. On the basis of the average score of these options, above mentioned were shortlisted as aspects. Targeted aspects are filtered and saved in CSV. Total number of reviews that were scraped are 4,886,406. The number of reviews after aspect level filtration reduced to 369,436 (lyrics = 7916, music

= 49238, song199248, video = 106127 and voice = 6907). Further, data of all files on one aspect is combined (for each aspect there is one file). Total files are 5, one for each aspect and named sub-dataset e.g., sub-dataset—voice.

3.2.2 Lowercasing

Several factors directly affect the analysis. It is observed that collected data contains the reviews in lowercase as well as uppercase text. For the same text (i.e., word) different patterns (mix of lower- and upper-case letters) were found like good, Good, GOOD, that type of data (both uppercase and lowercase) cases issues in classification, classifiers found different deviations in results. To avoid these issues all the text is converted in lowercase [44].

3.2.3 Noise Removal

The reviews that are scraped contains a lot of noise i.e., numbers, special characters and punctuations, that had nothing to do with SA. The extra data increases the computational cost of classifiers [45] as well as affects the classification results [46]. To get better results this noise is removed from the dataset.

3.2.4 Emojis Removal

Another popular way to express their views about anything is using Emoji's. It is easy way to show anyone feelings towards anything so users use it widely. Users leave their sentiment by posing appropriate emoji's [47]. Emojis are removed from the dataset because this research focuses text SA.

3.2.5 String Standardization

It is found in the data set that few reviews too long. Long string sizes condense the performance of the classifiers [48]. To overcome the performance problems, reviewing the reviews of each aspects, the maximum string length is defined for each Aspect. The aspect "lyrics" have a review in which the number of tokens is 11,487 that may affect the performance of the classifier. To resolve this issue the string size of "lyrics" is trimmed to 300 tokens that cover the 77.68% data. The aspect "music" has a comment in which the number of tokens is 9,914, the string size of "music" is defined maximum up to 150 characters that hold 81.28% of the total data. In "song" the string has a maximum number of tokens are 32,759. The string for "song" is defined as 150 tokens per instance that holds 77.67% of the total data. The aspect "video" string size is defined as 150 characters per example that covers 88.00% of the total dataset. In "video" the maximum size of the string is 10,510 tokens. The aspect "voice" has a review that has 32,759 tokens. The "voice" string size is trimmed up to 150 tokens that tackle 89.50% of the total data.

The aspect "lyrics" has 2,352,407 number of tokens before preprocessing, after applying preprocessing techniques the number of tokens was reduced to 1,226,830 tokens. Aspect "music" has 8,402,990 and 4,770,486 number of tokens before and after preprocessing respectively. The total number of tokens of aspect "song" before preprocessing is 15,962,658 and this number reduced up to 11,595,811 tokens. "video" has 12,639,362 tokens before preprocessing and after preprocessing this number reduced to 5,801,640. Number of tokens of aspect, "Voice", before and after processing are 4,269,028 and 672,323 respectively. Now the data is preprocessed and ready for the processing.

3.3 Corpus Generation

One of the two goals of this research is corpus generation. This section demonstrates the way it is generated for aspect level sentiment analysis. The corpus is named as CALSAS. Steps taken to build CALSAS are discussed in details in Sub-sections 4.1 and 4.2. Rest of the process is as follows:

3.3.1 Data Annotation

Data annotation is a static part and is a crucial component for any analysis. It is a way to categorize the object into one of the targeted classes, e.g., in this research there are three target classes—positive, negative and neutral, so every user review is labelled with one of these classes. The complete process of data annotation on this data set is discussed in [27].

3.3.2 Corpus Traits

The CALSAS consist of 369,436 reviews after all cut offs. Reviews are annotated into its targeted classes, one review assigned one of the target classes—negative, positive or neutral. Dataset is divided in five parts on the basis of aspects i.e., sub-dataset—lyrics, sub-dataset—music, sub-dataset—song, sub-dataset—video and sub-dataset—voice. 7,916 reviews belong to sub-dataset—lyrics, reviews 49,238 reviews belong to sub-dataset—music, sub-dataset—song consist of 199,248 reviews, sub-dataset—video have 106,127 reviews, sub-dataset—voice have total of 6,907 reviews. The entire corpus contains 24,534,221 number of tokens in 369,436 reviews. Detailed statics of the CALSAS can be seen in Tab. 2.

Table 2: Characteristics of CALSAS

Attribute	Value
Total reviews	369,436
Positive reviews	256,524
Negative reviews	36,764
Neutral reviews	76,151
Num. of reviews scraped for Corpus	4,886,406
Num. of tokens before preprocessing	43,626,445
After trim function num. of tokens	29,047,762
Review max. length before preprocessing	32,759
Review min. length before preprocessing	04
Review avg. length before preprocessing	113.05
Num. of tokens after preprocessing	24,534,221
Aspect lyrics number of reviews	7,916
Aspect music number of reviews	49,238
Aspect song number of reviews	199,248
Aspect video number of reviews	106,127
Aspect voice number of reviews	6,907
Review avg. length after preprocessing	66.41
Review max. length after preprocessing	300
Review min. length after preprocessing	04

4 Modeling

In this section, the experimental details are addressed which acquire to get the best results on the corpus CALSAS. To perform aspect level sentiment analysis, the experimentation was performed by using five different machine learning algorithms. Later on, the songs are rated on the basis on their aspects. The details of the experimental model are as follows. The CALSAS comprises 369,436 English reviews which belongs to five sub-datasets (i) sub-datasets—lyrics, (ii) sub-datasets—music, (iii) sub-datasets—song, (iv) sub-datasets—video and (v) sub-datasets—voice, collected from YouTube. Each review falls into one of three classes: positive, neutral and negative. As discussed above five different machine learning algorithms are employed. RapidMiner tool is used for the implementation of these algorithms. To validate the model, K-fold cross-validation technique is used with $K = 10$. Sentiment classification is majorly divided into two types—binary and polynomial. In binary, number of targeted classes are only two (positive and negative). In polynomial classification, number of target classes may be three or more. This research targets the polynomial type of English reviews [49]. Target classes, in this research, are three (i.e., positive, neutral and negative). Algorithms being employed, to design classification model, in this research, are NB, GB-Tree, KNN, LR) and ANN. The dataset splits into 8:2, to check the performance of the model on classification of reviews i.e., for the testing purpose, this study uses 80% of the total data and remaining 20% data is used for testing of the moel (While the data splitting, it is assured that every split contain data from all songs and classes, in balanced way). After getting the testing results, to validate the model performance model is validated with K-Fold cross-validation technique with $K = 10$. Four different model evaluation metrics—Precision, Recall, F-score and accuracy, are calculated (against each algorithm). The model performance is measured using accuracy metric.

4.1 Imbalanced Class Handling

The CALSAS is skewed towards positive class. Imbalanced data cause unreliable classification results [50]. To overcome the class imbalance issues different techniques are used e.g., Oversampling [51], Under-sampling [52] and Cluster-based oversampling [53]. In this study to handle class imbalance issue, oversampling is performed on three aspects i.e., Lyrics, Music and Voice, because these sub-datasets have a smaller number of records. While under-sampling is performed on rest of the two i.e., Video and Song, because these datasets are big enough. There exists different oversampling and under-sampling methods to balance the dataset. In oversampling different methods are used to balance the skewed dataset i.e., random oversampling, ADASYN (Adaptive Synthetic Sampling), smoting, borderline smooting, smoot-NC, Kmean smooting, and SVM smmoting etc [54–56]. For the under-sampling method to balanced the dataset there exist different techniques i.e., Random under-sampling for the majority class, NearMiss, Condensed Nearest Neighbor Rule, TomekLinks, Edited Nearest Neighbor Rule and Cluster Centroids etc. [57,58]. For the oversampling, random oversampling technique is implemented on three sub-datasets (Sub-dataset—lyrics, Sub-dataset—music and Sub-dataset—voice) to balance the dataset. The sub-datasets, Video and Song, have a large number of reviews so, under-sampling technique is implemented using Random under-sampling for the majority class method.

4.2 Results

There were three objectives of this research—dataset creation, classification and rating. Dataset creation process and its final outcome i.e., CALSAS, with its metadata has already been discussed in above sections. Subsequent sections will discuss rest of the two in detail.

4.2.1 Classification on Original CALSAS

Dataset CALSAS is comprised of five sub-datasets. Named after each aspect. For aspect level sentiment analysis and classification different Machine Learning Algorithms—NB, GB, KNN, LR and ANN, are tried for the best model design. On sub-dataset—lyrics, ANN and GB outperformed the other classifiers with 73.21% and 73.15% accuracies, where the value of F-score is 84.53% and 84.44% respectively. At the validation of the model ANN achieved 75.59% accuracy and 85.52% F-score which is the highest score. The validation accuracy of Gradient Boost on sub-dataset—lyrics is 70.69% and the value of F-score is 82.45%. On sub-dataset—music, Logistic Regression outperformed the other classifiers. It achieved 74.92% accuracy, where the F-score is 85.11% on test data. At cross validation LR achieved 74.73% accuracy with 84.94% F-score. ANN, on sub-dataset—music, achieved 71.70% accuracy with 83.50% F-score, at test dataset, whereas at cross validation, the model ANN achieved 74.73% accuracy with 84.94% F-score. On sub-dataset—song, Logistic Regression outperformed the other classifiers with 78.04% accuracy and 87.22% F-score. At cross validation of model, the accuracy of LR is 79.11% where the value of F-score is 87.77%. On sub-dataset—video, Logistic Regression outperformed the other classifiers with 59.96% testing accuracy and 74.26% cross validation accuracy. The value of F-score is 63.94% and 75.87% respectively. The performance of Gradient Boost Tree and ANN is also close to the LR. On sub-dataset—voice, ANN outperformed rest of the classifiers with 73.21% test accuracy and 67.97% cross validation accuracy. The value of F-score is 84.53% and 80.93% respectively. Detailed results can be seen in Fig. 1 and Tab. 3.

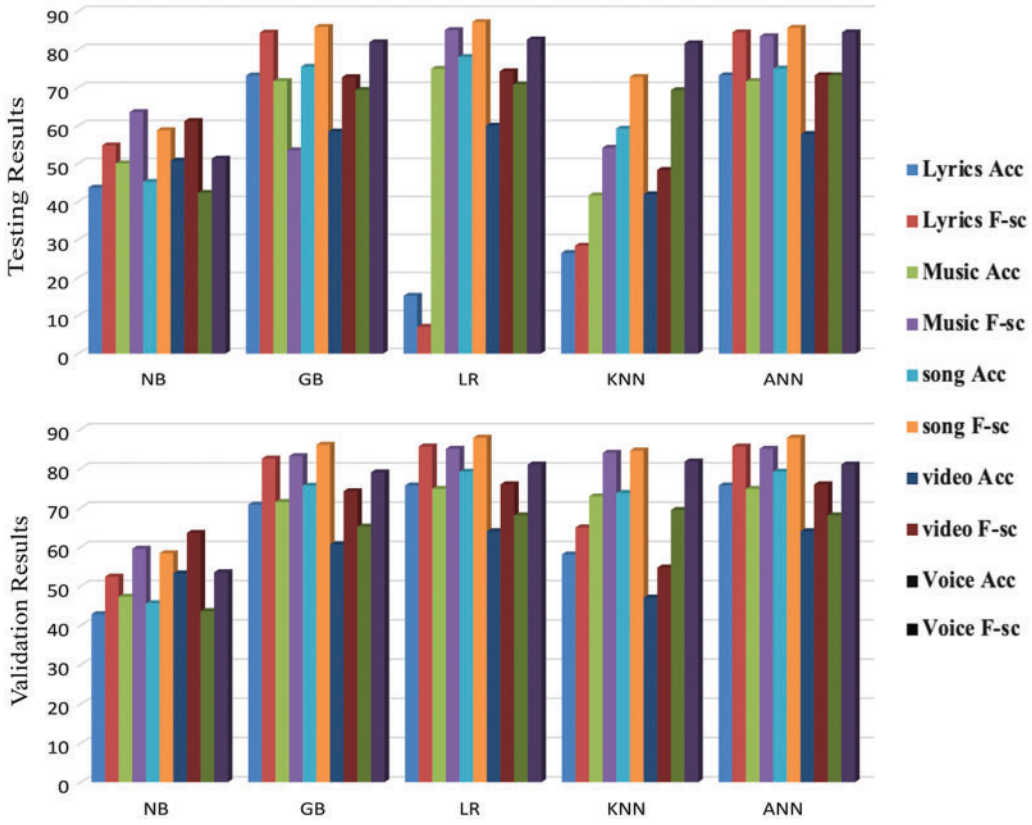

Figure 1: Comparison of testing & validation results

Overall weighted accuracy (OWA) on the whole dataset was calculated using Eq. (2).

Table 3: Testing and validation results before handling imbalance of dataset

Algo's	Evaluation metric	Naïve Bayes		Gradient Boost		Logistic Regression		KNN		ANN	
		Test	Vald.	Test	Vald.	Test	Vald.	Test	Vald.	Test	Vald.
Lyrics	**Accuracy**	43.67	42.80	73.15	70.69	15.22	75.59	26.47	58.01	**73.21**	**75.59**
	Recall	40.87	38.09	99.56	93.95	3.67	100	17.29	58.01	**100**	**100**
	Precision	83.13	83.81	73.31	73.46	**100**	**74.71**	79.52	73.69	73.21	74.71
	F-score	54.8	52.37	84.44	82.45	7.08	85.52	28.40	64.91	**84.53**	**85.52**
Music	**Accuracy**	50.04	47.24	71.7	71.39	**74.92**	**74.73**	41.6	72.80	71.70	74.73
	Recall	51.43	46.44	100	97.39	**100**	**98.85**	42.59	98.36	**100**	**98.85**
	Precision	**83.43**	**82.68**	71.70	72.41	74.08	74.46	74.36	73.17	71.70	74.46
	F-score	63.53	59.47	53.51	83.06	**85.11**	**84.94**	54.15	83.91	83.50	84.94
Song	**Accuracy**	45.20	45.57	75.47	75.54	**78.04**	**79.11**	59.19	73.67	74.99	79.11
	Recall	43.90	43.34	100	99.78	**100**	**100**	67.08	92.19	**100**	**100**
	Precision	**88.95**	**89.15**	75.36	75.52	77.35	78.21	79.56	78.01	74.99	78.21
	F-score	58.77	58.32	85.95	85.97	**87.22**	**87.77**	72.79	84.51	85.70	87.77
Video	**Accuracy**	50.76	53.20	58.43	60.59	**59.96**	**63.94**	41.91	47.02	57.79	63.94
	Recall	53.66	56.41	84.49	93.26	**100**	**93.96**	40.93	48.38	**100**	**93.96**
	Precision	**71.13**	**72.60**	63.85	61.50	59.07	63.62	59.01	62.87	57.78	63.62
	F-score	61.17	63.49	72.73	74.13	**74.26**	**75.87**	48.33	54.68	73.24	75.87
Voice	**Accuracy**	42.29	43.57	69.37	65.11	70.82	67.97	69.30	69.34	**73.21**	**67.97**
	Recall	37.06	39.27	**100**	**92.57**	99.90	92.32	92.38	93.36	100	92.32
	Precision	**83.33**	**83.75**	69.37	68.79	70.47	72.04	73.14	72.63	73.21	72.04
	F-score	51.30	53.47	81.91	78.93	82.64	80.93	81.64	81.70	**84.53**	**80.93**

$$OWA = \frac{\left(\sum_{a=1}^{|A|} EM_a \times N_a\right)}{\sum_{a=1}^{|A|} N_a} \tag{2}$$

where ***EM*** is Evaluation Metric

N is Number of text examples i.e., records

N_a is Number of text examples against aspect **a** i.e., records

|*A*| is count of aspects i.e., 5

On the basis of OWA, ANN outperformed rest of the algorithms with 70.18% test and 72.27% cross validation accuracies.

Classification on Updated Balanced Dataset (CALSASbalanced)

As discussed above the dataset CALSAS is skewed towards the positive class. The analysis using an imbalance dataset are not reliable. To get stable results the dataset CALSAS is balanced using the mix of oversampling and under-balancing.

For the oversampling, random oversampling technique is implemented on three sub-datasets (Lyrics, Music and Voice) to balance the dataset. The sub-datasets, Video and Song, have a large

number of reviews so, under-sampling technique is implemented using Random under-sampling for the majority class method.

4.2.2 Songs Rating

After resolving data imbalance issue same set of algorithms is re-run on the new dataset. In new results the ANN outperformed all the classifiers used in this research on the basis of accuracy and F-score. On sub-dataset—Lyrics, ANN outperformed the rest with 73.39% and 84.80%, accuracy and F-score respectively. On sub-dataset—Music, ANN outperformed the rest with 71.70% and 83.48%, accuracy and F-score respectively. On sub-dataset—Voice, ANN outperformed the rest with 69.30% and 81.51% accuracy and F-score, respectively. ANN outperformed the other classifiers, in all sub-datasets, after overpowering the class imbalance issue. At the undersampling of data the Gradient Boost Tree outperformed the other classifiers with 99.96% accuracy on both sub-datasets (song and video). The comparison of results on oversampling model is shown in Tab. 4 and Fig. 2.

Table 4: Results comparison after handling im-balancing of dataset

		NB	GB	LR	KNN	ANN
			Over-sampling			
Lyrics	Accuracy	41.88	25.77	15.22	26.47	**73.39**
	Recall	38.17	13.36	3.67	17.29	99.91
	Precision	83.08	82.7	100	79.52	73.66
	F-score	52.3	23.15	7.08	28.4	84.8
Music	Accuracy	51.36	36.31	28.31	47.6	**71.7**
	Recall	53.97	37.17	21.7	42.59	100
	Precision	82.97	76.46	100	74.36	71.7
	F-score	65.39	50.02	35.66	54.15	83.48
Voice	Accuracy	42.14	39.83	20.85	**69.3**	53.95
	Recall	36.85	34.97	5.01	92.38	64.09
	Precision	83.25	81.91	100	73.14	73.71
	F-score	51.08	49.01	9.54	81.51	68.56
			Under-sampling			
Song	Accuracy	52.92	**99.96**	52.41	53.84	46.77
	Recall	37.57	99.97	45.97	9.07	38.67
	Precision	73.49	100	58.29	100	52.38
	F-score	49.71	99.97	51.4	16.63	44.49
Video	Accuracy	49.37	**99.96**	33.09	34.95	33.43
	Recall	40.94	100	0	2.55	100
	Precision	49.56	99.93	0	100	33.43
	F-score	44.84	99.96	0	4.98	50.11

Songs are rated based on their aspects. This research targets the five aspects of songs, based upon people's opinion, i.e., Lyrics, Music, Song, Video and Voice. Two kinds of ratings are generated. First is aspect level i.e., rating of the song on the basis of one aspect. Second is overall rating of the song based upon the polarity of the comments.

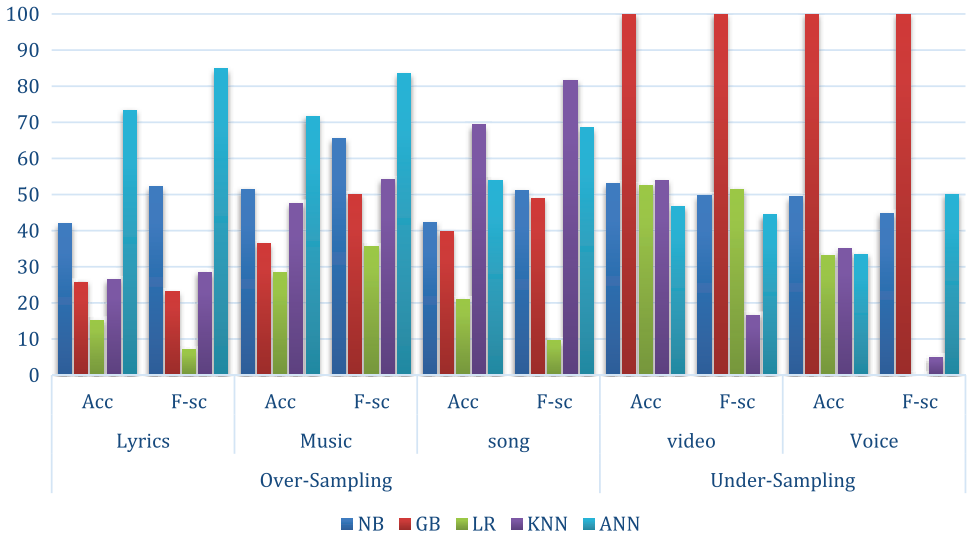

Figure 2: Results comparison after handling im-balancing of dataset

To calculate Aspect Level Rating (ASR), first of all mean polarity of each aspect of each song is calculated using Eq. (3). The rating can be generated by sorting these values of mean polarity (MP) of one aspect on the basis of all songs and highest rated can be find out by applying Max operator given in Eq. (4).

$$MP_{a,s} = \frac{\left(\sum_{i=1}^{N_{a,s}} P_{a,s}\right)}{N_{a,s}} \tag{3}$$

$$\text{ALR}_a = \max\left(\forall_{s \in S} MP_{a,s}\right) \tag{4}$$

where

$\boldsymbol{s} \in \boldsymbol{S}$ *and* $\boldsymbol{S}$ *is set of songs*

$\boldsymbol{a} \in \boldsymbol{A}$ *and* $\boldsymbol{A}$ *is set of Aspacts*

$\boldsymbol{MP_a}$ *is mean polarity of one aspect* $\boldsymbol{a}$

$\boldsymbol{N_a}$ *is Number of text examples against aspect* $\boldsymbol{a}$ *i.e. records*

$\boldsymbol{P_a}$ *is polarity of one text record i.e. one review*

$\boldsymbol{N_{a,s}}$ *is Number of text records against aspect* $\boldsymbol{a}$ *and song* $\boldsymbol{s}$

$\boldsymbol{P_{a,s}}$ *is polarity of one text record against aspect* $\boldsymbol{a}$ *and song* $\boldsymbol{s}$

Eq. (4) will return the name of the song with highest polarity in that aspect a. Repeating it for all aspects can generate best songs for each aspect. In experiments this process is repeated for all, five, aspects. These songs are rated highest among all others based upon one aspect. Details of songs' rating at five different aspects is shown in Tab. 5.

Table 5: Songs rating; from highest to lowest in each aspect

Rate	Lyrics	Music	Song	Video	Voice	Overall
1	Criminal	5-Sugar	5-Sugar	5-Sugar	Bailando	Roar
2	5-Sugar	Shape of You	Bailando	Roar	Shape of You	Shape of You
3	Shake It Off	Bailando	Roar	Shake It Off	Roar	5-Sugar
4	Roar	Roar	Criminal	Shape of You	Chantaje	Shake It Off
5	Shape of You	Shake It Off	Shape of You	This Is What You Came For	Sorry	Bailando f
6	Bailando	Criminal	Chantaje	Criminal	Love The Way You Lie	Criminal
7	Sorry	This Is What You Came For	Sorry	Love The Way You Lie	Shake It Off	Sorry
8	Love The Way You Lie	Chantaje	Shake It Off	Sorry	This Is What You Came For	Love The Way You Lie
9	Chantaje	Love The Way You Lie	Love The Way You Lie	Bailando	Criminal	Chantaje
10	This Is What You Came For	Sorry	This Is What You Came For	Chantaje	5-Sugar	This Is What You Came For

The set that was returned was $\{Criminal, 5-Suger, 5-Suger, 5-Suger, Bailandoft.DescemerBueno\}$ i.e., name of the song with top ranking for the set of $\{Lyrics, Music, Song, Video, Voice\}$.

To find the best overall song we just need to take the mean polarity of all songs and highest can be reported as overall best song. Eq. (5) serve the purpose.

$$\text{OSR} = \left(\max\left(\frac{\forall_{y\in S}\left(\sum_{i=1}^{N} P\right)}{N}\right)\right) \tag{5}$$

Katy Perry's Song, *Roar*, is at top with highest overall polarity.

5 Conclusions

This research is carried to perform Aspect Level Sentiment Analysis of reviews of songs. A benchmark dataset of English text is presented. For benchmark dataset, reviews from 10 top rated songs on YouTube, are scrapped. A short review was conducted to shortlist the aspect to study. Dataset is organized as a set of five sub-datasets named Music, Lyrics, Song, Voice and Video. After pre-processing that include different filtrations specially aspect level filtration. After pre-processing data was annotated using automated channel. Five commonly known Machine Learning Algorithms—Naïve Bayes, Gradient Boost Tree, Logistic Regression, K-Nearest Neighbors and Artificial Neural Network is applied to perform Aspect Level Sentiment Analysis. Main experiment was performed twice, once on original dataset *CALSAS* (with imbalance classes) and then the same experiment was

repeated after handling the class imbalance issue. The new dataset was named *CALSASbalanced*. In first experimentation, for Lyrics and Voice, ANN proved to be the best with 73.21% and 73.21% accuracy, respectively. For Aspects, Music, Song and Video, Logistic Regression outperformed the rest of the algorithms with 74.92%, 78.04% and 59.96% accuracy, respectively. In second experimentation, on *CALSASbalanced*, for Lyrics and Music, ANN proved to be the best with 73.39% and 71.70% accuracy, respectively. For Aspect, Voice, KNN outperformed the rest of the algorithms with 69.30% accuracy. For Aspects, Song and Video, Gradient Boost outperformed the rest of the algorithms with same 99.96% accuracy. The model is validated with K-fold cross-validation technique with $k = 10$ and no alarming underfitting or overfitting is witnessed. ANN outperformed the other ML Algorithms, implemented in this study, on classification, validation, and oversampling. So, this study recommends ANN for the Text classification of reviews, especially on large datasets. At the end, song's rating on aspect level as well as overall are calculated. According to the overall rating, the song *Roar* is rated as a top song. At Aspect Level, the songs named *Criminal, 5-Suger, 5-Suger, 5-Suger, Bailando ft.Descemer Bueno* beat their counterparts on the basis of the aspects *Lyrics, Music, Song, Video and Voice*, respectively.

Acknowledgement: The authors would like to express their most profound gratitude towards, Mr. Muneeb Fazal and Mr. Burhan Ul Haq Zahir for their valuable time and efforts for helping us in data collection and in the annotation process.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. A. Mostafa and M. Z. Saringatb, "Comparative analysis for arabic sentiment classification," in *First Int. Conf. on Applied Computing to Support Industry: Innovation and Technology*, Ramadi, Iraq, 1174, pp. 271–285, 2020

[2] A. Madden, I. Ruthven and D. M. Menemy, "A classification scheme for content analyses of youtube video comments," *Journal of Documentation*, vol. 69, no. 5, pp. 693–714, 2013.

[3] Z. Papacharissi, "Sentiment analysis of roman urdu/hindi using supervised methods," *Ain Shams Engineering Journal*, vol. 2, no. 3, pp. 1093–1113, 2013.

[4] M. A. Qureshi, M. Asif, M. F. Hassan, A. Abid, A. Kamal *et al.,* "Sentiment analysis of reviews in natural language: Roman urdu as a case study," *IEEE Access*, vol. 10, pp. 24945–24954, 2022.

[5] H. L. Vogel, "The virtual sphere 2.0: The internet, the public sphere and beyond," in *IEEE Int. Conf. on Fuzzy Systems*, Hyderabad, India, vol. 2, pp. 164–172, 2013.

[6] L. Jiang, C. Li, S. Wang and L. Zhang, "Deep feature weighting for naive bayes and its application to text classification," *Engineering Applications of Artificial Intelligence*, vol. 52, no. 7, pp. 26–39, 2016.

[7] J. P. Verma, B. Patel and A. Patel, "Big data analysis: Recommendation system with hadoop framework," in *Proc. 2015 IEEE Int. Conf. on Computational Intelligence and Communication Technology*, Ghaziaabad, India, pp. 92–97, 2015.

[8] M. A. C. Jondar, "Rich youtuber, poor Youtuber: Implementasi business intelligence dalam meningkatkan pendapatan channel youtube ye," Undergraduate thesis, University of Surabaya, 2020.

[9] S. Moon, P. K. Bergey and D. Lacobucci, "Dynamic effects among movie ratings, movie revenues and viewer satisfaction," *Journal of Marketing*, vol. 74, no. 1, pp. 108–121, 2010.

[10] S. Zhang, T. Aktas and J. Luo, "Mi youtube es su youtube? Analyzing the cultures using youtube thumbnails of popular videos," in *IEEE Int. Conf. on Big Data*, Singapore, pp. 4999–5006, 2020.

[11] S. Choi and A. Segev, "Finding informative comments for video viewing," *SN Computer Science*, vol. 1, no. 1, pp. 47–60, 2020.
[12] D. J. Kalita, V. P. Singh and V. Kumar, "A survey on svm hyper-parameters optimization techniques," in *Social Networking and Computational Intelligence*, Berlin, Germany, Springer, pp. 243–256, 2020.
[13] S. Badugu, "Telugu movie review sentiment analysis using natural language processing approach," in *Data Engineering and Communication Technology*. Berlin, Germany, Springer, pp. 685–695, 2020.
[14] P. K. Mallick, V. E. Balas, A. K. Bhoi and A. F. Zobaa, *Congnitive Inforamtics and Soft Computing*. vol. 768. Singapore: Springer, 2019.
[15] E. Guresh, N. Fink, D. Friesenhahn, N. Ramkumar and M. J. F. Gerald, "Techniques for managing persistent document collections," *Google Patents*, U.S. Patetnt No. 9626362, Washington, DC, U.S.A, 2019.
[16] D. Patel, S. Shah and H. Chhinkaniwala, "Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique," *Expert Systems with Applications*, vol. 134, pp. 167–177, 2019.
[17] D. V. Thin, L. S. Le, H. M. Nguyen and N. L. T. Nguyen, "A joint multi-task architecture for document-level aspect-level sentiment analysis in vietnamese," *International Journal of Machine Learning and Computing*, vol. 12, no. 4, pp. 126–135, 2022.
[18] P. Patil and P. Yalagi, "Sentiment analysis using aspect level classification," *Acadenia*, vol. 4, no. 4, pp. 23–27, 2016.
[19] H. Wang and Y. Wang, "A review of online product reviews," *Journal of Service Science and Management*, vol. 13, no. 1, pp. 88–96, 2020.
[20] H. H. Do, P. W. C. Prasad, A. Maag and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: A comparative review," *Expert Systems with Applications*, vol. 118, no. 6, pp. 272–299, 2019.
[21] J. Philip, A. Baby and A. Kannammal, "The good, the bad and ugely: Openion mining analysis on user tweets in twitter," *Journal of Emerging Technologies and Innovative Research*, vol. 6, no. 5, pp. 1–7, 2019.
[22] K. Sarawgi and V. Pathak, "Opinion mining: Aspect level sentiment analysis using sentiwordnet and amazon web services," *International Journal of Computer Applications*, vol. 158, no. 6, pp. 31–36, 2017.
[23] Y. Wang, M. Huang, A. Sun and X. Zhu, "Aspect-level sentiment analysis using as-capsules," in *Proc. of World Wide Web Conf.*, San Francisco, pp. 2033–2044, 2019.
[24] R. Kumar, H. S. Pannu and A. K. Malhi, "Aspect-based sentiment analysis using deep networks and stochastic optimization," *Neural Computing Applications*, vol. 32, no. 8, pp. 3221–3235, 2019.
[25] B. G. Patra, D. Das and S. Bandyopadhyay, "Multimodal mood classification of hindi and western songs," *Journal of Intelligent Information Systems*, vol. 51, no. 3, pp. 579–596, 2018.
[26] F. Chen, Z. Yuan and Y. Huang, "Multi-source data fusion for aspect-level sentiment classification," *Knowledge-Based Systems*, vol. 187, no. 1–2, pp. 104831–104838, 2020.
[27] M. A. Qureshi, M. Asif, M. F. Hassan, G. Mustafa, M. K. Ehsan *et al.,* "A novel auto-annotation technique for aspect level sentiment analysis," *Computers, Materials & Continua*, vol. 70, no. 3, pp. 4987–5004, 2022.
[28] V. Gupta, V. K. Singh, P. Mukhija and U. Ghose, "Aspect-based sentiment analysis of mobile reviews," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 4721–4730, 2019.
[29] S. Kiritchenko, X. Zhu, C. Cherry and S. M. Mohammad, "Nrc-canada-2014: Detecting aspects and sentiment in customer reviews," in *Proc. of 8th Int. Workshop on Sementic Evaluation*, Dublin, Ireland, pp. 437–442, 2014.
[30] C. Brun, D. N. Popa, C. Roux and D. Maupertuis, "Xrce: Hybrid classification for aspect-based sentiment analysis," in *Proc. of 8th Int. Workshop on Sementic Evaluation*, Dublin, Ireland, pp. 838–842, 2014.
[31] M. Syahrul and M. Dwi, "Aspect-based sentiment analysis to review products using naïve bayes," in *AIP Conf. Proceedings*, Poland, 1867, pp. 20060–20068, 2017.
[32] P. Gamallo, M. Garcia and C. L. Technology, "Citius: A naive-bayes strategy for sentiment analysis on english tweets," in *Proc. of 8th Int. Workshop on Sementic Evaluation*, Dublin, Ireland, pp. 171–175, 2014.
[33] S. Vanaja, "Aspect-level sentiment analysis on e-commerce data," in *Int. Conf. on Inventive Research in Computing Applications*, Coimbatore, India, pp. 1275–1279, 2018.

[34] M. Afzaal, M. Usman and A. Fong, "Predictive aspect-based sentiment classification of online tourist reviews," *Journal of Information Science*, vol. 45, no. 3, pp. 341–364, 2018.
[35] S. C. Sekharan, "Aspect based sentiment analysis of amazon product reviews," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 22, pp. 151–157, 2018.
[36] A. Nawaz, A. Ahmed, A. Tariq, A. Muhammad and R. Rashid, "Product's behaviour recommendations using free text: An aspect based sentiment analysis approach," *Cluster Computing*, vol. 1, pp. 1267–1279, 2019.
[37] S. Reddy, "Aspect based sentiment analysis of students opinion using machine learning techniques," in *Int. Conf. on Inventive Computing and Informatics*, Coimbatore, India, pp. 726–731, 2017.
[38] O. Alqaryouti, N. Siyam, A. Abdel and K. Shaalan, "Applied computing and informatics aspect-based sentiment analysis using smart government review data," *Applied Computing and Informatics*, vol. 16, pp. 1–20, 2019.
[39] R. S. Jagdale, V. S. Shirsat and S. N. Deshmukh, *Sentiment Analysis on Product Reviews Using Machine Learning Techniques*. Singapore: Springer, 2019.
[40] P. Pandey and N. Soni, "Sentiment analysis on customer feedback data: Amazon product reviews," in *Int. Conf. on Machine Learning, Big Data, Cloud and Parallel Computing*, Faridabad, India, pp. 320–322, 2019.
[41] K. F. Frasincar and R. Dekker, "An information gain-driven feature study for aspect-based sentiment analysis," in *Int. Conf. on Applications of Natural Language to Information Systems*, Salford, United Kinkdom, pp. 48–59, 2016.
[42] L. Xu, J. Liu, L. Wang and C. Yin, "Aspect based sentiment analysis for online reviews," in *Advances in Computer Science and Ubiquitous Computing*. vol. 2. Singapore: Springer, pp. 24–32, 2017.
[43] Y. W. Wu and B. P. Bailey, "Better feedback from nicer people: Narrative empathy and ingroup framing improve feedback exchange," *ACM Human-Computer Interaction*, vol. 4, no. CSCW3, pp. 1–20, 2021.
[44] M. Asif, M. A. Qureshi, A. Abid and A. Kamal, "A dataset for the sentiment analysis of indo-pak music industry," in *Int. Conf. on Innovative Computing*, Lahore, Pakistan, pp. 1–6, 2019.
[45] A. Z. Syed, M. Aslam and A. M. M. Enriquez, "Lexicon based sentiment analysis of urdu text using sentiunits," in *Maxican Int. Conf. on Artificial Intelligence*, Berlin, Springer, pp. 32–43, 2010.
[46] G. Qi, Z. Zhu, K. Erqinhu, Y. Chen, Y. Chai *et al.,* "Fault-diagnosis for reciprocating compressors using big data and machine learning," *Simulation Modeling Practice and Theory*, vol. 80, no. 11, pp. 104–127, 2018.
[47] Y. Elazar and Y. Goldberg, "Adversarial removal of demographic attributes from text data," in *Conf. on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 11–21, 2018.
[48] J. Hartmann, J. Huppertz, C. Schamp and M. Heitmann, "Comparing automated text classification methods," *International Journal of Research in Marketing*, vol. 36, no. 1, pp. 20–38, 2019.
[49] Z. Mahmood, I. Safdar, R. M. A. Nawab, F. Bukhari, R. Nawaz *et al.,* "Deep sentiments in roman urdu text using recurrent convolutional neural network model," *Information Processing and Management*, vol. 57, no. 4, pp. 102233–102246, 2020.
[50] V. López, A. Fernández and F. Herrera, "On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed," *Information Sciences*, vol. 257, no. 2–3, pp. 1–13, 2014.
[51] R. Gillala, V. K. Reddy and A. K. Tyagi, "Kdos: Kernel density based over sampling:- A solution to skewed class distribution," *Journal of Information Assurance & Security*, vol. 15, no. 2, pp. 40–52, 2020.
[52] D. Devi, S. K. Biswas and B. Purkayastha, "A review on solution to class imbalance problem: Undersampling approaches," in *Int. Conf. on Computational Performance Evaluation*, Shillong, India, pp. 626–631, 2020.
[53] G. Rekha, V. K. Reddy and A. K. Tyagi, "A novel approach for solving skewed classification problem using cluster based ensemble method," *Mathematical Foundations of Computing*, vol. 3, no. 1, pp. 1–9, 2020.
[54] S. Wang, Z. Li, W. Chao and Q. Cao, "Applying adaptive over-sampling technique based on data density and cost-sensitive SVM to imbalanced learning," in *Int. Joint Conf. on Neural Networks*, Brisbane, QLD, Australia, pp. 1–8, 2012.

[55] P. Wibowo and C. Fatichah, "Pruning-based oversampling technique with smoothed bootstrap resampling for imbalanced clinical dataset of COVID-19," *Journal of King Saud Universit-Computer and Information Sciences*, pp. 1–10, In press, 2021.
[56] J. V. F. França, J. M. C. D. Santos, F. Henrique, V. Garcia, C. Manfredini *et al.,* "Legal judgment prediction in the context of energy market using gradient boosting," in *IEEE Int. Conf. on Systems, Man, and Cybernetics*, Oronto, Canada, pp. 875–880, 2020.
[57] H. Wang and X. Liu, "Undersampling bankruptcy prediction: Taiwan bankruptcy data," *PLoS One*, vol. 16, no. 7, pp. 1–17, 2021.
[58] Y. S. Jeon and D. -J. Lim, "Psu: Particle stacking undersampling method for highly imbalanced big data," *IEEE Access*, vol. 8, pp. 131920–131927, 2020.