Computers, Materials & Continua
DOI:10.32604/cmc.2022.020544
Article

Tech Science Press

A Novel Auto-Annotation Technique for Aspect Level Sentiment Analysis

Muhammad Aasim Qureshi[1,*], Muhammad Asif[1], Mohd Fadzil Hassan[2], Ghulam Mustafa[1], Muhammad Khurram Ehsan[1], Aasim Ali[1] and Unaza Sajid[1]

[1]Department of Computer Sciences, Bahria University, Lahore Campus, 54000, Pakistan
[2]Computer and Information Science Department, University Teknologi, Petronas, 32610, Malaysia
[*]Corresponding Author: Muhammad Aasim Qureshi. Email: maasimq@hotmail.com
Received: 27 May 2021; Accepted: 28 June 2021

Abstract: In machine learning, sentiment analysis is a technique to find and analyze the sentiments hidden in the text. For sentiment analysis, annotated data is a basic requirement. Generally, this data is manually annotated. Manual annotation is time consuming, costly and laborious process. To overcome these resource constraints this research has proposed a fully automated annotation technique for aspect level sentiment analysis. Dataset is created from the reviews of ten most popular songs on YouTube. Reviews of five aspects—voice, video, music, lyrics and song, are extracted. An N-Gram based technique is proposed. Complete dataset consists of 369436 reviews that took 173.53 s to annotate using the proposed technique while this dataset might have taken approximately 2.07 million seconds (575 h) if it was annotated manually. For the validation of the proposed technique, a sub-dataset—Voice, is annotated manually as well as with the proposed technique. Cohen's Kappa statistics is used to evaluate the degree of agreement between the two annotations. The high Kappa value (i.e., 0.9571%) shows the high level of agreement between the two. This validates that the quality of annotation of the proposed technique is as good as manual annotation even with far less computational cost. This research also contributes in consolidating the guidelines for the manual annotation process.

Keywords: Machine learning; natural language processing; annotation; semi-annotated technique; reviews annotation; text annotation; corpus annotation

1 Introduction

In recent years, the internet has gained popularity and it has become an eminent platform for socializing among users [1]. It has transformed the real world into a cyber-world [2]. Now almost everyone has easy access to hand-held devices with a reliable internet connection. Over the last few years, it is witnessed that due to the popularity of these handheld gadgets, massive data is being generated on the daily basis [3]. This bulk data is being generated from diverse sources like social media platforms, e-commerce, games, etc [4]. The generated data is both in a structured and unstructured form [3]. Most of the unstructured data is being produced by e-users (i.e., people using Twitter, WhatsApp, Facebook, Instagram, YouTube, etc.) [4]. This unstructured data is the prime challenge for the data analysts.

To overcome this challenge there exist pre-processing techniques like data cleaning, dimensionality reduction, data standardization, data transformation and data annotation etc. that structure the unstructured data. Data annotation is one of the preprocessing techniques [5,6]. It is a process to label the data into its targeted classes [7,8]. There exist different schemes of data annotation like fully-automated annotation [9], manual annotation [10] and semi-automated annotation [11]. Certain applications and platforms use an annotation scheme that suits them best based on their customized requirements. Manual annotation is a way to declare the subjectivity of each entity present in the data according to the metadata file i.e., annotation guidelines by involving humans [12]. It is considered ideally reliable and accurate than the other two schemes. At the same time, it requires time, cost and efforts, which makes it less practical. A fully-automated annotation scheme is a way to annotate the documents with the help of fully automated tools like Portable document format annotation (PDFAnno), MyMiner, BARAT rapid annotation tool (BRAT) and team-text annotation tool (TeamTat) etc. A semi-automated annotation scheme is a way to annotate the dataset both taken manual and fully-automated annotation schemes into the consideration.

Availability of different social networks, online blogs and other forums that enable people to discuss various aspects of products or services. Using new technologies, most social networking or e-commerce websites allow users to express their experience regarding the products, services and features [13]. These reviews can help in analyzing any product, service and company [14].

The exponential growth of reviews/comments can be witnessed due to the drastic increase in the number of e-users [15]. The internet has re-modelled the communication world as the backbone of a digital era [16]. Now the showbiz industry is also using this paradigm to progress [17]. Due to its accessibility and innovativeness, people can, now, easily access the entertainment contents, watch them and give their feedback in the form of likes and reviews. Further, this content is judged by these likes, rating and reviews on it [18]. A simple formula to check the quality of the content is:

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

$$where,\ \ CQ = TL - TD$$

where CQ quantifies the content quality using the metrics of Total likes (TL) and Total dislikes (TD).

This provides limited insight into the quality of the content. A better way to evaluate the quality of content is through the analysis of comments/reviews. If the count of these reviews is not so high, this goal can easily be achieved by reading and analyzing these reviews manually. It becomes humanely impossible to analyze these reviews if they are in huge amount. That creates a need to analyze these reviews through a proper automated channel. In this context Sentiment analysis (SA), as an important paradigm, is used to know the general opinion towards that content [19]. SA is a way to categorize the people's opinions towards the entity into positive, neutral or negative [20]. It can also be said that it is a way to classify the sentiments according to the class assigned by the reviewer [21].

As discussed above, a huge amount of unstructured data is being generated by e-users on daily basis [22] in the form of comments and reviews. To analyze this data and mine the hidden patterns, the data is required to be in a structured form. There exist different preprocessing [23] techniques to overcome different data anomalies and prepare them for analysis. In Data Annotation every entity of the dataset is assigned a label according to their subjectivity. There exist different semi-automated and automated tools to annotate different types of data contents like the video [12], audio [24], image [11] and text [25].

The rest of the paper is organized into six sections. Section 2 focuses on the previous related research on annotation. Section 3 discusses the entire corpus generation process including the steps involved. Section 4 focuses on the N-gram based proposed technique of auto-annotation for the English text data. Section 5 discusses the experimental results obtained from the proposed technique. Finally, Section 6 concludes the paper.

2 Literature Review

State-of-the-art studies have presented different tools for annotation. These tools can be classified into two categories that are annotation for image data and annotation for text data.

2.1 Annotation for Image Data

Bio-notate is a web-based annotation tool for biomedical annotation to annotate gene-disease and to annotate the binary association between the proteins [26]. AlvisAE, reported in [27], is a semi-automated annotator to annotate tasks and assign different rules, based on the expertise and generate automatic annotation which can also be modified by the users. It is mostly used in biology and crop-sciences. GATE teamware [28] is a web-based open-source semi-automatic annotator which performs pre-annotation of fungal enzymes with the facility of manual correction.

2.2 Annotation for Text Data

Catma [29], is a web-based annotator which allows the users to import the text data using document browsing as well as by using Hypertext markup language (HTML) document by entering Uniform resource locater (URL). It allows the corpus creation. It also has the capability of automated document annotation as well as assigning manual tag sets. FLAT [30] folia is a web-based annotator which provides linguistic and semantic-based annotation using Folia format to annotate the biomedical document. MAT [31] is an active learning tool to annotate the text by importing a file in Extensible markup language (XML) and exports the annotations in either XML or Javascript object notation (JSON) formats. It is an offline application to annotate the text. BRAT, reported in [32], is a web-based text annotation tool to support Natural language processing (NLP) such as named entity recognition and part of speech tagging. BioQRator [33] is another web-based tool to annotate biomedical literature.

TeamTat [34] presents an open-source web-based document annotation tool that annotates the plain text inputs as well as document input (BioC XML or XML) and the output document is in the form of BioC XML inline annotated document. Djangology [35] is a collaborative based document annotator to annotate the documents using web services. To annotate, the document is imported as plain text and after annotation, the annotated document is exported in plain text format. In [36] geo-annotator is presented. It is a collaborative semi-automated platform for constructing geo-annotated text corpora. The annotator is a semi-automatic web-based tool with collaborative visual analytics to solve place references in Natural language.

There exist some articles annotator, Loomp [37] was a web-based tool or the annotation of articles that annotate articles based on the article‘s annotator. RDFa [38] based on a general-purpose annotation framework, to annotate the news articles automatically. MyMiner [39] a web-based annotation tool can retrieve the abstracts and create a corpus for the annotation.

A plain document is imported to find a binary relationship or for tagging the entity and output is also exported in the plain text. WebAnno [40] provides full functionality for syntax and semantic-based annotations. It allows a variety of formats to import the document for the annotation and as well to export the annotated document. PDFAnno [41] a PDF document annotator available open-source. The document is imported to annotate in PDF file format the PDFAnno performs annotation and to find the relationship between entities. It can also provide the facility to annotate the figures and tables as well. The tagtog [42] provides annotation at the entity level and as well as document level. It uses an active learning approach to annotate the retrieved abstracts or full test retrieve for the annotation purpose. LightTag [43] is a commercial tool to annotate the text and its supports different languages. It can learn from active annotators using machine learning and annotate the unseen text.

There exist different automated tools to annotate the image data and text data as well. BRAT [32] is a tool that performs intuitive annotation, named entity annotation and dependency annotation. Where ezTag [44] tool is used to annotate the medical-based text data using lexicon base tagging concepts. CAT [45] is a tool that annotates the Ribonucleic acid (RNA) sequences and annotates the clades and to identify the relationship in orthology. According to the best of our knowledge, there hardly exists any tool to annotate the English text (comments/reviews) for SA.

Recent studies like [46–48], have witnessed that researchers are annotating the text manually and some of them by using TextBlob [49–52]. There exist tools like PDFAnno [41], MyMiner [39], BtableRAT [32] for text annotation, but no literature has witnessed text annotation for sentiment analysis at the aspect level. Manual annotation of reviews is a very hectic and time taking task [12] e.g., this research has figured out that on average 5.6 s are required to annotate one review.

This study presents a corpus of 369,436 reviews, annotated through N-gram based proposed technique. If the manual annotation were performed, it might have taken 2.07 million seconds (574.68 h) i.e., approximately 24 days to annotate. Manual text annotation is the bottleneck in NLP because it is very time consuming [12]. To overcome this bottleneck, this study presents an automated annotation technique for the English text using N-Gram based technique at the aspect level. The technique is also validated with Cohen's Kappa Coefficient value. After the validation of the technique, the entire corpus is annotated at the aspect level using the N-gram based proposed technique.

3 Corpus Generation

A quality corpus needs systematic collection and thorough preprocessing which can further be divided into three sub-tasks named data collection, preprocessing and data annotation. Details can be seen below:

3.1 Dataset Collection

Data is a vital part of any analysis. No analysis can be performed without data. To collect data and to build a gold-standard dataset, the top ten songs are selected [50]. Details can be seen in Tab. 1.

3.2 Preprocessing

Data quality directly affects data analysis [18]. To separate reviews carrying targeted aspects and to get processed data, different preprocessing techniques are applied like aspect filtration, data integration, lowercasing, emojis' removal and string size standardization. Details are as below.

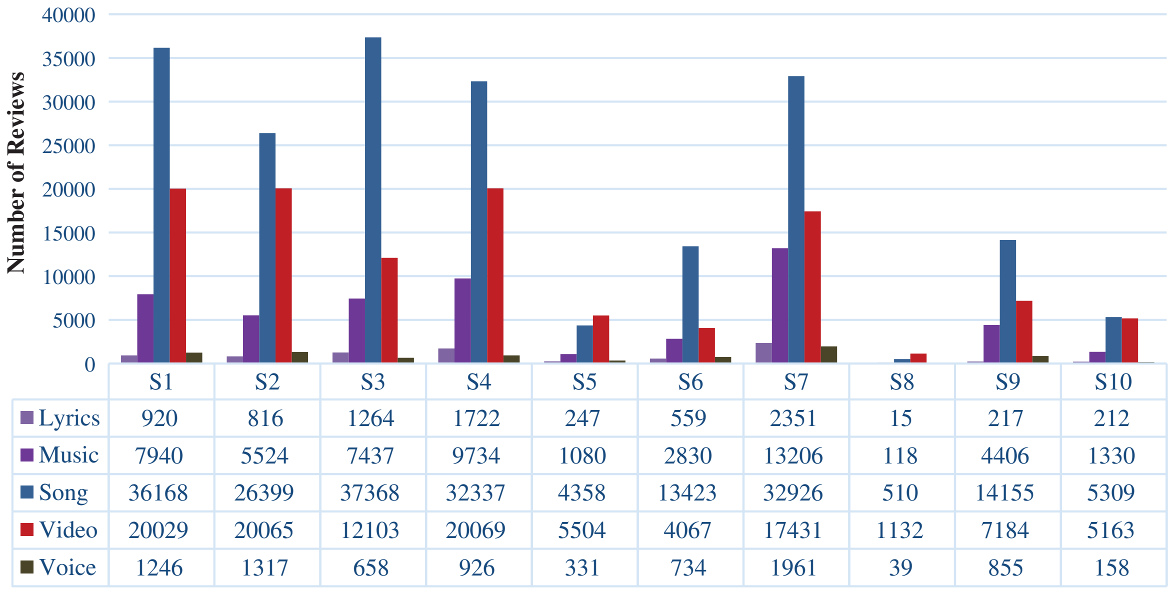
Table 1: Songs and number of reviews scraped

No	Singer	Song Title	Views	Reviews
S1	Justin Bieber	Sorry	3,203,542,747	816,063
S2	Katy Perry	Roar	2,932,210,456	638,490
S3	Ed Sheeran	Shape of You	4,449,412,646	910,006
S4	Taylor Swift	Shake It Off	2,832,020,062	517,249
S5	Shakira	Chantaje	2,455,461,250	380,955
S6	Rihanna	Calvin Harris-This Is What You Came	2,275,349,039	287,379
S7	Eminem	Love The Way You Lie ft. Rihanna	1,868,708,629	523,115
S8	Natti Natasha	Ozuna Criminal	1,889,274,387	259,644
S9	Maroon 5	Sugar	3,047,850,854	342,142
S10	Enrique Iglesias	Bailando ft. Descemer Bueno, Gente De Zona	2,774,918,985	211,363

3.2.1 Aspect Filtration

In this study, five aspects/features (lyrics, music, song, video and voice) are targeted for auto-annotation of reviews. Reviews that contain these aspects are separated by applying filters and saved in CSV file format. Total 4,886,406 reviews are scraped and after aspect level filtration, the obtained number of record is 369,436.

The pre-processing extracted 7916 reviews for lyrics, 49238 for music, 199248 for the song, 106127 for video and 6907 for voice. Dataset is now in fifty data files, details can be viewed in Fig. 1.



	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
■ Lyrics	920	816	1264	1722	247	559	2351	15	217	212
■ Music	7940	5524	7437	9734	1080	2830	13206	118	4406	1330
■ Song	36168	26399	37368	32337	4358	13423	32926	510	14155	5309
■ Video	20029	20065	12103	20069	5504	4067	17431	1132	7184	5163
■ Voice	1246	1317	658	926	331	734	1961	39	855	158

Figure 1: Number of reviews per song per aspect after aspect level filtration

3.2.2 Data Integration

The data from ten different files of one aspect are gathered in one file. As this study covers the five different aspects which resulted overall five files, one for each aspect and we call them sub-dataset. Each sub-dataset is named after one aspect e.g., sub-dataset—Voice.

3.2.3 Lowercasing

The case though has no special impact on the analysis of the data but when the same data is presented in different cases then it has adverse effects [48] e.g., algorithms will consider "Yes" and "yes" two different values. Therefore, to overcome this effect, whole data is converted into lowercase.

3.2.4 Noise Removal

It is reported, time and again, that noise directly affects the classification results [51]. It was noted that collected data contained a lot of noise like white spaces, special characters, punctuation signs, etc. that has nothing to do with the analysis. To improve the quality of data all these characters were removed.

3.2.5 Remove Number

The dataset contained English text and numbers as well. This study is to analyze English text. The extra data increase computational power and also diverse the results [49]. To address the said problems, all numbers are removed.

3.2.6 Remove the Emoji's

Emoji's is a popular way to show one's feelings. It is widely used by e-users to show their feelings towards the entity. They leave their sentiments using emoji's [52]. This study only focused on the text, therefore, emoji's are removed from the dataset.

3.2.7 Trim String Size

In the dataset, there were several reviews of extraordinary length. For example in sub-dataset—Lyrics, a review had 11,487 number of tokens. In the same way in sub-dataset—Music there was a review that had 9,914 tokens. Such lengthy reviews are outliers and have a bad impact on classification [53]. To overcome the issue length standardization is applied. To improve the quality of data, to have the least impact of data loss maximum length size is defined as 150 for all sub-datasets except for lyrics (due to the very small number of reviews).

To resolve this issue the string size of sub-dataset—Lyrics is trimmed to 300 tokens that cover the 77.68% data and for the rest of the sub-datasets the max string size is defined as 150 characters. The rest of the details can be seen in Tab. 2.

Table 2: Trimming ratio

Aspect	Lyrics	Video	Music	Voice	Song
String size (in characters)	300	150	150	150	150
Captured data	77.68%	81.28%	77.67%	88.00%	89.50%

Average reduction in tokens in sub-datasets—Lyrics, Music, Song, Video and Voice is 47.85%, 43.23%, 27.36%, 54.10% and 84.25% respectively. Details of tokens before and after preprocessing can be seen in Fig. 2.

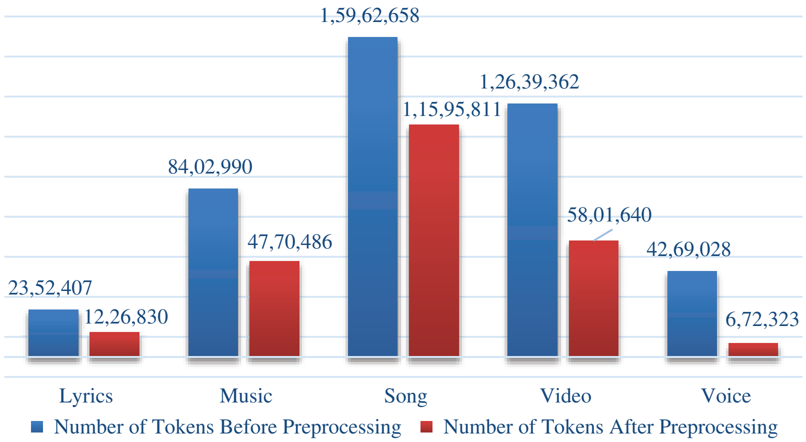


Figure 2: Number of tokens, before and after preprocessing

3.3 Data Annotation

Being the static part, data annotation is a process of categorizing the text (e.g., instance, review or comment) into positive, neutral or negative based upon its subjectivity. Previous studies showed different ways to annotate the data i.e., auto-annotation, semi-annotation and as well as manual annotation. Many automated tools are witnessed to annotate the image and video data.

Very few tools exist to annotate text data, especially, there is no automated tool exists to annotate the English text for sentiment analysis. Generally, manual annotation is used to label text data. Details can be seen in the subsequent section.

3.3.1 Manual Annotation

In manual annotation, each review is labelled according to its subjectivity, manually. Each review is labelled by reading it one by one and assigned class according to its behaviour as positive, neutral or negative. Following the process explained in [54] manual annotation process was divided into four steps.

In step-I, to annotate the reviews manually, the guidelines are prepared. In step-II, three volunteers were contacted to annotate the data. To start with they were given basic training based upon the guidelines. In step-III, the dataset was given to the annotators for annotation. The conflicts were resolved using an inter-annotator agreement. Finally, in step-IV, the computational value of the inter-annotator agreement was calculated using Kappa-statistics. The details of all steps are as below:

Annotation Guidelines Preparation

In the light of guidelines for each class—positive, negative and neutral, presented in different research works [55–57] are mapped on the current problem. Details are as below:

Guidelines for Positive Class

A review will be assigned "Positive"

- If it shows positive sentiments [53].
- If its behaviour is both neutral and positive [53,58].
- If there exists some positive word(s) in the sentence [59] e.g., good, beautiful, etc.
- If there exist illocutionary speech act like wow, congrats and smash classified as positive [60].

Examples: In "best music yet" the word "best" clearly shows the positive polarity of the aspect—music. In another review "his voice it's so soft and cool," the behaviour is positive towards aspect—voice.

Guidelines for Negative Class

A review will be assigned "Negative"

- If it shows negative sentiments [55].
    - ○ If the use of language is abusive [1].
    - ○ If the behaviour is un-softened [59].
- If there exist some negative word(s) in the sentence [60] e.g., bad, ugly, annoying, etc.
- If there exist negation, in a review e.g., not good.

Examples: In "music is trash," the word trash expresses the polarity i.e., negative of the review for the aspect—music. In, "this is the stupid voice," the word "stupid" shows negative sentiments of the reviewer on aspect—voice.

Guidelines for Neutral Class

A review will be assigned "neutral"

- If it is not showing any positive or negative sentiments [56].
- If a review has a piece of realistic information [57].
- If a review has both positive and negative sentiments [61].

Examples: In "that music going too far away" the subjectivity of the review isn't clear so it will be annotated as neutral. In, "the video is neither good nor bad," in the review both sentiments are present so it will be annotated as neutral.

Training of Annotators

For the manual annotation the help of three volunteers was pursued, let's call them A, B and C. The volunteers were graduates, well familiar with reviews and concepts of annotation and had a good grip on the English language. Three hours of the hands-on training session was conducted to explain the guidelines and discuss possible issues with them.

Conflict Resolution

For conflict resolutions, a short sample dataset of 100 reviews was created let's call it SSD100. SSD100 was given to the first two annotators—annotator A and annotator B. Once they completed the annotation, a short meeting was arranged to resolve the conflicts by involving the third annotator too. After the conflict resolution, SSD100 was given to the annotator C for annotation.

3.3.2 Problems Faced During Manual Annotation

Though volunteers were very cooperative but still the process faced few problems as listed below:

(i) Training
(ii) Individual's perception

(iii) Clash removals
(iv) Confidence level

Even after 3 h of hands-on practice, still, annotators were consulting the trainers for the resolution of the issues. There was a big issue of an individual's perception. 5.33% of manual annotations, done by three annotators, was still updated by the trainer. During the annotation process, they were also asked to mention their confidence level (1–10) regarding their annotated label. The average confidence was 90.50%. It took almost 6 h to annotate 3700 reviews with an average of 5.6 s per review. This showed that manual annotation even with qualified and trained annotators is not perfect and unseen and unreported lags always remain there. A sample of annotated data is shown in Tab. 3.

Table 3: Sample annotated data

Reviews	Class
very melodious voice	Positive
I love justin bieber voice	Positive
your voice is so cute	Positive
superb voice i like it	Positive
even their voices	Neutral
is it me or is this voice about selena	Neutral
why is this my first time hearing his voice	Neutral
hello everyone is sut voice here	Neutral
his voice	Neutral
im addicted of his voice	Negative
her voice so annoying	Negative
the voice is addictive	Negative
this is stupid voice	Negative

4 Proposed Technique for Auto-Annotation

To overcome the hectic and time-consuming process of manual annotation, this study has is presented a new fully automated technique for text annotation (at aspect level) based upon the language modelling technique of N-gram.

4.1 N-Gram

Models that assign probabilities to sequences of words are called language models (LMs). N-gram is one of the simplest models that assign probabilities to sentences and sequences of words. An N-gram is a sequence of N words. For example in a sentence "Best song ever justin…" a 2-gram (or bigram) is a two-word sequence like "Best song," "song ever," or "ever justin," and a 3-gram (or trigram) is a three-word sequence of words like "Best song ever," or "song ever justin." N-gram model estimates the probability of the last word of an n-gram given the previous words, and also assign probabilities to entire sequences, thus the term N-gram is used to mean either the word sequence itself or the predictive model that assigns it a probability.

For the joint probability of each word in a sequence having a particular value $P(W = w_1,\ X = w_2,\ Y = w_3,\ \ldots; Z = w_n)$ we'll use $P(w_1, w_2, w_3, \ldots, w_n)$.

Applying the chain rule to the words

$$\mathrm{P}(w_1^n) = \mathrm{P}(w_1)\mathrm{P}(w_2 \mid w_1)\mathrm{P}(w_3 \mid w_1^2)\ldots\mathrm{P}(w_n \mid w_1^{n-1}) = \prod_{k=1}^{n}\mathrm{P}(w_k \mid w_1^{k-1}) \tag{2}$$

where w_1^n *is* $w_1, w_2, w_3, \ldots, w_n$ and $P(w_x \mid w_1^y)$ is conditional probability of occurrence of w_x given the occurrence of $w_1, w_2, w_3, \ldots, w_y$.

For Bi-gram i.e., N = 2 Eq. (2) can be updated as

$$\mathrm{P}(w_1^n) = \mathrm{P}(w_1)\mathrm{P}(w_2 \mid w_1)\mathrm{P}(w_3 \mid w_2)\ldots\mathrm{P}(w_a \mid w_{a-1})\ldots\mathrm{P}(w_n \mid w_{n-1}) = \prod_{k=1}^{n}\mathrm{P}(w_k \mid w_{k-1}) \tag{3}$$

4.2 Re-Definition of Ps

This research is redefining *P*'s to solve the problem in hand. We need to find the polarity of the text having n-words i.e., $w_1, w_2, w_3, \ldots, w_n$, we define it as $\mathrm{P}(w_1^n)$. The polarity will be counted if the behavior of the word is with respect to the aspect. So the polarity is being checked in the form of pair of words, out of which one is supposed to be aspect. $\mathrm{P}(w_x \mid w_y)$ defines the two words polarity where w_x is aspect and w_y can be anything. If w_y is positive then the polarity of these two words' combination would be positive and if w_y is negative then the polarity of these two words combination would be negative else neutral.

The polarity of two-word combination is required to be checked for all occurrences of w_y *where* $1 \leq y \leq n$ *and* $y \neq x$. To find the aspect in all words of the text w_x is varied from 1 to n. Hence we have Eq. (2) to express all this where:

$$\begin{aligned} P(w_1^n) = {} & P(w_1)P(w_1 \mid w_2)P(w_1 \mid w_3)\ldots P(w_1 \mid w_n) \\ & .P(w_2)P(w_2 \mid w_1)P(w_2 \mid w_3)\ldots P(w_2 \mid w_n) \\ & .P(w_3)P(w_3 \mid w_1)P(w_3 \mid w_2)\ldots P(w_3 \mid w_n) \\ & \ldots \\ & \ldots \\ & .P(w_n)P(w_n \mid w_1)P(w_n \mid w_2)\ldots P(w_n \mid w_{n-1}) \end{aligned} \tag{4}$$

$$= \prod_{k=1}^{n} P(w_k)P(w_k \mid w_1)P(w_k \mid w_2)\ldots P(w_k \mid w_{n-1}) \tag{5}$$

$$= \prod_{k=1}^{n}\prod_{\substack{i=1 \\ i \neq k}}^{k-1} P(w_k \mid w_1^{k-1}) \tag{6}$$

where w_1^n *is* $w_1,\ w_2,\ w_3, \ldots,\ w_n$

where w_k *is the* ***aspect***

$P(w_x)$ *is the occurance of aspect*

$\mathrm{P}(w_x \mid w_y)$ is the polarity of occurrence of aspect, w_x, given the occurrence of w_y

We define the value of P as below:

$$P(w_x) = \begin{cases} p & \mathit{if}\ w_x \in Bag_p \\ n & \mathit{if}\ w_x \in Bag_p \\ 1 & \mathit{if}\ w_x \notin Bag_p\ \mathit{and}\ Bag_n \end{cases} \tag{7}$$

And

$$\mathrm{P}(w_a \mid w_{a-1}) = \begin{cases} p, & \mathit{if}\ \ w_a = Aspect\ \ and\ \ w_{a-1}\ \in Bag_p \\ n, & \mathit{if}\ \ w_a = Aspect\ \ and\ \ w_{a-1}\ \in Bag_n \\ p, & \mathit{if}\ \ w_a \neq Aspect\ \ and\ \ \mathit{if}\ \ w_a\ \ and\ \ w_{a-1} \in\ \ Bag_p \\ n, & \mathit{if}\ \ w_a \neq Aspect\ \ and\ \ \mathit{if}\ \ w_a \in\ \ Bag_p\ \ and\ \ w_{a-1} \in Bag_n \\ n, & \mathit{if}\ \ w_a \neq Aspect\ \ and\ \ \mathit{if}\ \ w_a\ \ or\ \ w_{a-1} \in\ \ Bag_n \\ 1, & otherwise \end{cases} \tag{8}$$

where Bag_p and Bag_n, are bag of positive words and bag of negative words respectively (the list of these words can easily be found online (e.g., GitHub) that can then be updated according to the tokens).

For a complete sentence, this will result in an expression like

$$p^x n^y \quad where\ 0 \le x, y \le n \tag{9}$$

Irrespective of the powers, p^x is assigned 1 and n^y is assigned -1

$$\mathrm{P}(w_n \mid w_{n-1}) = p^x n^y = \begin{cases} 1, & label = positive \\ -1, & label = negative \end{cases} \tag{10}$$

For example, if the value of $\mathrm{P}(w_n \mid w_{n-1}) = p^3 n^2$ we will assign $p^3 = 1\ \mathit{and}\ n^2 = -1$.

In this way, nearest words are associated with the targeted aspect and after computing its value, a label is assigned. To understand, Eq. (8) is explained, condition by condition, through examples as below:

Example: $\boldsymbol{p}$, $\boldsymbol{if\ \ w_a = Aspect\ \ and\ \ w_{a-1} \in Bag_p}$

The value p will be assigned to $\mathrm{P}(w_a \mid w_{a-1})$, if w_a is an aspect and w_{a-1} is a word that belongs to the bag of positive words. E.g., in "*justin bieber you have amazing voice*" if we check the value of $\mathrm{P}(w_a \mid w_{a-1})$ at highlighted words then it will be p as $w_a = Voice$, which is an aspect and $w_{a-1} = amazing$ which belongs to bag of positive words.

Example; $\boldsymbol{n}$, $\boldsymbol{if\ \ w_a = Aspect\ \ and\ \ w_{a-1} \in Bag_n}$

The value n will be assigned to $\mathrm{P}(w_a \mid w_{a-1})$, if w_a is an aspect and w_{a-1} is a word that belongs to the bag of negative words. E.g., in "*justin bieber you have annoying voice*" if we check the value of $\mathrm{P}(w_a \mid w_{a-1})$ at highlighted words then it will be n as $w_a = Voice$, which is an aspect and $w_{a-1} = annoying$ which belongs to bag of negative words.

Example: $\boldsymbol{p}$, $\boldsymbol{if\ \ w_a \neq Aspect\ \ and\ \ if\ \ w_a\ \ and\ \ w_{a-1} \in Bag_p}$

The value p will be assigned to to $\mathrm{P}(w_a \mid w_{a-1})$, if w_a is not an aspect and w_a and w_{a-1} is a word that belongs to the bag of positive words. E.g., in "*love justin bieber voice look is so amazing*" if we check the value of $\mathrm{P}(w_a \mid w_{a-1})$ at highlighted words then it will be p as $w_a \neq Voice$, which is not an aspect and w_a & w_{a-1} belongs to bag of positive words.

Example: $\boldsymbol{n}$, $\boldsymbol{if\ \ w_a \neq Aspect\ \ and\ \ if\ \ w_a \in Bag_p\ \ and\ \ w_{a-1} \in Bag_n}$

The value n will be assigned to $\mathrm{P}(w_a \mid w_{a-1})$, if w_a is not an aspect and w_a is a word that belongs to the bag of negative words and w_{a-1} is a word that belongs to the bag of negative words. E.g., in "*ad to say but justin biebers voice is lighter not good as baby old justin bieber*" if we check the w_a that belongs to the bag of positive words where w_{a-1} is belongs to bag of negative words.

Example: $\boldsymbol{n}$, $\boldsymbol{if\ \ w_a \neq Aspect\ \ and\ \ if\ \ w_a\ \ or\ \ w_{a-1} \in Bag_n}$

The value n will be assigned to $\mathrm{P}(w_a \mid w_{a-1})$, if w_a is not an aspect and w_a and w_{a-1} is a word that belongs to the bag of negative words. E.g., in "*three billion viewers of justin hates i want to hear his voice*" if we check the value of $\mathrm{P}(w_a \mid w_{a-1})$ at highlighted words 'hates' then it will be n as $w_a \neq Voice$, which is not an aspect and w_a & w_{a-1} belongs to bag of negative words.

Example: 1, ***&otherwise***

If the review not having any word that belongs to the bag of positive nor from the bag of negative words, the value 1 is assigned to that type of reviews which is labelled as neutral. E.g., in "*the october from bangladesh who are with me rise your voice*" not any single word that belongs to the bag of positive words or belongs to the bag of negative words, the polarity of all these types of reviews is declared as neutral.

5 Validation of Proposed Technique

The technique is validated using Cohen's Kappa statistics. It is a statistic that is used to indicate the inter-rater reliability between the annotators [62]. According to Cohen's Kappa, the value of Kappa >90% indicates almost perfect agreement between the annotators [63]. Tab. 4 presents the details of the interpretation of different levels of values of Cohen's Kappa.

Table 4: Interpretation of cohen's kappa

Value of Kappa	Level of agreement
40%–59%	Weak
60%–79%	Moderate
80%–90%	Strong
Above 90%	Almost perfect

To prove the efficacy of the proposed technique, the inter-annotator agreement is calculated using Cohen's Kappa statistics. For this purpose, three experiments were conducted. Two on SSD100 and one on sub-dataset—Voice. The details of the experiments are shown in Tab. 5.

Table 5: Comparison of kappa statistics value

Dataset	Dataset size	Annotation type	Annotation type	Kappa statistics
SSD100	100	Manual (three annotators)		0.852843%
SSD100	100	Manual	Auto	0.909699%
Voice	6907	Manual	Auto	0.957139%

5.1 Experiment 1

In this experiment, SSD100 was used for the manual annotation. This experiment was conducted during the training of Annotators—A, B and annotator C. Kappa value i.e., inter-annotator agreement between the three annotators was calculated (the value appeared 85.28%).

5.2 Experiment 2

In this experiment, the annotation results of SSD100 using manual annotation and using the based proposed technique are compared and the Kappa Statistics value is calculated. The estimated value of Kappa is 90.96%. The level of agreement using the value of Kappa are shown in Tab. 4.

5.3 Experiment 3

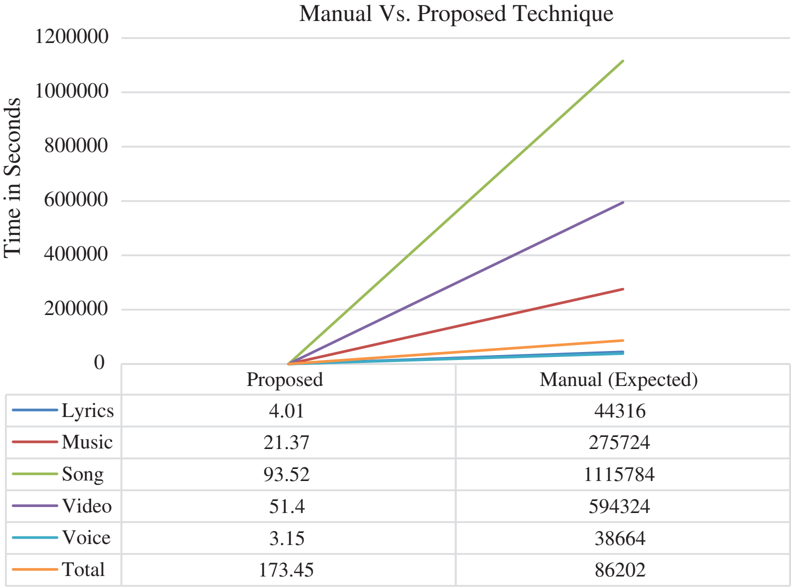
In this experiment the sub-dataset—Voice is annotated once with manual annotation technique and then using the proposed technique. Kappa Statistics value is calculated to validate the reliability of the results of the proposed technique. The value appeared to be 95.71%. This proves that the proposed technique is giving results as good as manual annotation even with far less computational cost. The details of the Kappa value of different sizes of datasets can be seen in Tab. 5.

The remaining sub-datasets of four aspects are also annotated using N-gram based proposed technique. Details of all sub-datasets can be seen in Tab. 6.

Table 6: Number of reviews regarding each aspect

Sub-dataset	Size
Lyrics	7916
Music	48238
Song	199248
Video	106127
Voice	6907
Total	369436

It took almost 2 min 53.45 s in annotating complete dataset of size 369,436 reviews of 24534205 tokens with an average of 0.46963 milliseconds per review and 7.07 microseconds per token If it was attempted manually then it had taken approximately 14 weeks, 1 day and 7 h while working 40 h a week with the average of 5.6 s per review and 23.44 microseconds per token. Fig. 3 explains the difference between the two. The Proposed technique completes the task in few seconds the manual might had taken days and weeks. In terms of the expected time of manual annotation, the proposed technique is efficient with a ratio of 1:11934.28.



	Proposed	Manual (Expected)
Lyrics	4.01	44316
Music	21.37	275724
Song	93.52	1115784
Video	51.4	594324
Voice	3.15	38664
Total	173.45	86202

Figure 3: Comperison of time between manual and proposed technique

6 Conclusion

This study also established that manual annotation is too subjective. It is not as good as it is supposed to be. Unknowingly inaccuracies do exist. This research has presented a new technique to annotate large datasets as good as manual annotation and even with far less computational cost in comparison to manual annotation. The dataset of English text reviews is scraped, preprocessed and annotated manually as well as using the proposed technique. This technique may benefit in multiple ways. It does not need additional resources—financial as well as human. It is very efficient and requires very less time without any additional cost. The performance ratio of manual to proposed i.e., 1:11934.28 shows its efficiency. If the complete dataset were annotated manually then it might had have taken approximately 2.07 million seconds, 14 weeks, 1 day and 7 h while working 40 h a week (i.e., 575 h) but this technique has done the same with 173.53 s (2 min and 53.45 s). The high value of Kappa statistics i.e., 95.71% shows and validates the reliability of results generated by proposed technique. Machine learning and deep learning algorithms can be applied to the datasets, one with manual annotation and other with this proposed technique, to extend this analysis and study the variation in the two techniques.

Acknowledgement: The authors would like to express their most profound gratitude towards, Mr. Umar Shoukat, Mr. Muneeb Fazal, Mr. Burhan Ul Haq Zahir and Ms. Rabbia Abrar 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 *Applied Computing to Support Industry: Innovation and Technology: First Int. Conf. ACRIT 2019*, Ramadi, Iraq, September 15–16, 2019, Revised Selected Papers, vol. 1174, pp. 271–285, 2020.

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

[3] S. Jain, D. Proserpio, G. Quattrone and D. Quercia, "Nowcasting gentrification using airbnb data," *Proceedings of the ACM on Human-Computer Interaction, CSCW*, NewYork, USA, vol. 5, pp. 1–20, 2021.

[4] D. Antonakaki, P. Fragopoulou and S. Ioannidis, "A survey of twitter research: Data model, graph structure, sentiment analysis and attacks," *Expert Systems with Applications*, vol. 164, pp. 114006–114027, 2021.

[5] Z. Ke, J. Sheng, Z. Li, W. Silamu and Q. Guo, "Knowledge-guided sentiment analysis via learning from natural language explanations," *IEEE Access*, vol. 9, pp. 3570–3578, 2021.

[6] P. Mylonas, Y. Voutos and A. Sofou, "A collaborative pilot platform for data annotation and enrichment in viticulture," *Information*, vol. 10, no. 4, pp. 149–175, 2019.

[7] E. S. Jo and T. Gebru, "Lessons from archives: strategies for collecting sociocultural data in machine learning," in *Proc. of the 2020 Conf. on Fairness, Accountability, and Transparency*, New York, USA, pp. 306–316, 2020.

[8] K. Yordanova, "Towards automated generation of semantic annotation for activity recognition problems," in *2020 IEEE Int. Conf. on Pervasive Computing and Communications Workshops*, PerCom Workshops 2020, Rostock, Germany, pp. 1–6, 2020.

[9] D. Kalita, "Supervised and unsupervised document classification-a survey," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 2, pp. 1971–1974, 2015.

[10] O. Grljević and Z. Bošnjak, "Sentiment analysis of customer data," *Strategic Management*, vol. 23, no. 3, pp. 38–49, 2018.

[11] O. Beijbom, P. J. Edmunds, C. Roelfsema, J. Smith, D. I. Kline *et al.*, "Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation," *PLOS One*, vol. 10, no. 7, pp. 1–22, 2015.

[12] M. Neves and J. Ševa, "An extensive review of tools for manual annotation of documents," *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 146–163, 2019.

[13] W. P. M. Wong, M. C. Lo and T. Ramayah, "The effects of technology acceptance factors on customer e-loyalty and e-satisfaction in Malaysia," *International Journal of Business and Society*, vol. 15, no. 3, pp. 477–502, 2014.

[14] 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, pp. 26–39, 2016.

[15] 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.

[16] P. Bhutani and A. Saha, "Towards an evolved information food chain of world wide web and taxonomy of semantic web mining," in *Int. Conf. on Innovative Computing and Communications*, Singapore, vol. 56, pp. 443–451, 2019.

[17] G. Negro, "How Chinese people use the internet," in *The Internet in China*, London, United Kingdom: Springer, pp. 89–142, 2017.

[18] Y. W. Wu and B. P. Bailey, "Better feedback from nicer people: Narrative empathy and ingroup framing improve feedback exchange," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW3, pp. 1–20, 2021.

[19] 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.

[20] D. M. Koupaei, T. Song, K. S. Cetin and J. Im, “An assessment of opinions and perceptions of smart thermostats using aspect-based sentiment analysis of online reviews,” *Building Environment*, vol. 170, pp. 106603–106614, 2020.

[21] P. Chiranjeevi, D. Teja Santosh and B. Vishnuvardhan, “Survey on sentiment analysis methods for reputation evaluation,” *Proceeding of Cognitive Informatics and Soft Computing*, 2017, vol. 768, pp. 53–66, 2019.

[22] E. W. T. Ngai, L. Xiu and D. C. K. Chau, “Application of data mining techniques in customer relationship management: A literature review and classification,” *Expert System Applpplications*, vol. 36, no. 2 PART 2, pp. 2592–2602, 2009.

[23] S. Vijayarani, M. J. Ilamathi and M. Nithya, “Preprocessing techniques for text mining—An overview,” *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[24] 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, Dec. 2018.

[25] I. Gupta, B. D. Eugenio, B. Ziebart, A. Baiju, B. Liu *et al.*, “Human-human health coaching via text messages: Corpus, annotation, and analysis,” in *Proc. of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Singapore, pp. 246–256, 2020.

[26] C. Cano, T. Monaghan, A. Blanco, D. P. Wall and L. Peshkin, “Collaborative text-annotation resource for disease-centered relation extraction from biomedical text,” *Journal of Biomedical Informatics*, vol. 42, no. 5, pp. 967–977, 2009.

[27] F. Papazian, R. Bossy and C. Nédellec, “AlvisAE: A collaborative web text annotation editor for knowledge acquisition,” in *Proc. of the Sixth Linguistic Annotation Workshop*, Jeju, Republic of Korea, pp. 149–152, 2012.

[28] K. B. Kalina, H. Cunningham, I. Roberts, A. Roberts, V. Tablan *et al.*, “GATE teamware: A web-based, collaborative text annotation framework,” *Language Resources and Evaluation*, vol. 47, no. 4, pp. 1007–1029, 2013.

[29] J. C. Meister, E. Gius, J. Horstmann, J. Jacke and M. Petris, “CATMA 5.0 tutorial,” *Digital Humanities*, vol. 1, pp. 1–2, 2017.

[30] M. V. Gompel and M. Reynaert, “Folia: A practical XML format for linguistic annotation—A descriptive and comparative study,” *Computational Linguistics in the Netherlands Journal*, vol. 3, pp. 63–81, 2013.

[31] R. Ai and M. Charfuelan, “MAT: A tool for L2 pronunciation errors annotation,” in *Proc. of the 9th Int. Conf. on Language Resources and Evaluation*, Reykjavik, Iceland, pp. 3979–3982, 2014.

[32] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou *et al.*, “BRAT: A web-based tool for NLP-assisted text annotation,” in *Proc. of the Demonstrations at the 13th Conf. of the European Chapter of the Association for Computational Linguistics*, Berlin, Germany, pp. 102–107, 2012.

[33] D. Kwon, S. Kim, S. -Y. Shin, A. Chatr-aryamontri and W. J. Wilbur, “Assisting manual literature curation for protein—Protein interactions using BioQRator,” *Database*, vol. 2014, pp. 1–7, 2014.

[34] R. Islamaj, D. Kwon, S. Kim and Z. Lu, “Teamtat: A collaborative text annotation tool,” *Nucleic Acids Research*, vol. 48, no. W1, pp. W5–W11, 2020.

[35] E. Apostolova, S. Neilan, G. An, N. Tomuro and S. Lytinen, “Djangology: A light-weight web-based tool for distributed collaborative text annotation,” *International Journal of Geo-Informatio*, vol. 8, no. 4, pp. 161–190, 2010.

[36] M. Karimzadeh and A. M. MacEachren, “Geoannotator: A collaborative semi-automatic platform for constructing geo-annotated text corpora,” *ISPRS Intternational Journal of Geo-Information*, vol. 8, no. 4, pp. 161, 2019.

[37] M. Luczak-Rösch and R. Heese, “Linked data authoring for non-experts,” *Linked Data on the Web Workshop*, Madrid, *Spain*, pp. 1–5, 2009.

[38] A. Khalili, S. Auer and D. Hladky, "The rdfa content editor-from wysiwyg to wysiwym," in *IEEE 36th Annual Computer Software and Applications Conf.*, Izmir, Turkey, pp. 531–540, 2012.

[39] D. Salgado, M. Krallinger, M. Depaule, E. Drula, A. V. Tendulkar *et al.*, "Myminer: A web application for computer-assisted biocuration and text annotation," *Bioinformatics*, vol. 28, no. 17, pp. 2285–2287, 2012.

[40] S. M. Yimam, I. Gurevych, R. E. de Castilho and C. Biemann, "Webanno: A flexible, web-based and visually supported system for distributed annotations," in *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Germany, pp. 1–6, 2013.

[41] H. Shindo, Y. Munesada and Y. Matsumoto, "Pdfanno: A web-based linguistic annotation tool for pdf documents," in *Proc. of the Eleventh Int. Conf. on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, pp. 1–5, 2018.

[42] J. M. Cejuela, P. McQuilton, L. Ponting, S. J. Marygold, R. Stefancsik *et al.*, "Tagtog: Interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles," *Database*, vol. 2014, pp. 1–8, 2014.

[43] J. M. Davis III, J. D. Davis and M. W. Kitchen, "Light-tag system." *Google Patents*, vol. 1, pp. 12–17, 2016.

[44] D. Kwon, S. Kim, C.-H. Wei, R. Leaman and Z. Lu, "Eztag: Tagging biomedical concepts via interactive learning," *Nucleic Acids Research*, vol. 46, no. W1, pp. W523–W529, 2018.

[45] I. T. Fiddes, J. Armstrong, M. Diekhans, S. Nachtweide, Z. N. Kronenberg *et al.*, *Comparative Annotation Toolkit (CAT)-Simultaneous Clade and Personal Genome Annotation*, New York, United States: Cold Spring Harbor Laboratory Press, pp. 1–10, 2017.

[46] M. Asif, M. A. Qureshi, A. Abid and A. Kamal, "A aataset for the sentiment analysis of indo-pak music industry," in *Int. Conf. on Innovative Computing*, Lahore, Pakistan, pp. 1–6, 2019.

[47] Z. U. Rehman and I. S. Bajwa, "Lexicon-based sentiment analysis for urdu language," in *6th Int. Conf. on Innovative Computing Technology*, Dublin, Ireland, pp. 497–501, 2017.

[48] A. A. Chaudhri, S. S. Saranya and S. Dubey, "A survey on analyzing covid-19 vaccines on twitter dataset using tweepy and textblob," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 3, pp. 8579–8581, 2021.

[49] Y. Zahidi, Y. El Younoussi and Y. Al-amrani, "Different valuable tools for arabic sentiment analysis: A comparative evaluation," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 1, pp. 753–762, 2021.

[50] Kworb, "kwrob.net," 2019. [Online]. Available: https://kworb.net/youtube/archive.html (Accessed 01 September 2019).

[51] G. Qi, Z. Zhu, K. Erqinhu, Y. Chen, Y. Chai and J. Sun, "Fault-diagnosis for reciprocating compressors using big data and machine learning," *Simulation Modelling Practice and Theory*, vol. 80, pp. 104–127, Jan. 2018.

[52] Y. Elazar and Y. Goldberg, "Adversarial removal of demographic attributes from text data," in *Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 11–21, 2018.

[53] 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.

[54] 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.

[55] S. Khedkar and S. Shinde, "Deep learning-based approach to classify praises or complaints from customer reviews," in *Proc. of Int. Conf. on Computational Science and Applications*, Pune, India, pp. 391–402, 2020.

[56] H. Hope, "Hello [Streamer] pogchamp': The language variety on twitch," University of Stavanger, Norway, Thesis, 2019.

[57] Z. Chen and T. Qian, "Transfer capsule network for aspect level sentiment classification," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, Italy, pp. 547–556, 2019.

[58] T. I. Jain and D. Nemade, "Recognizing contextual polarity in phrase-level sentiment analysis," *International Journal of Computer Applications*, vol. 7, no. 5, pp. 12–21, 2010.

[59] M. Abdul-Mageed and M. T. Diab, "AWATIF: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis," in *8th Int. Conf. on Language Resources and Evaluation*, Istanbul, Turkey, vol. 515, pp. 3907–3914, 2012.

[60] M. Ahmed, Q. Chen and Z. Li, "Constructing domain-dependent sentiment dictionary for sentiment analysis," *Neural Computing and Applications*, vol. 32, pp. 1–14, 2020.

[61] M. D. P. S. Zárate, J. M. Moreira, K. L. Ortiz, H. L. Aveiga, M. Á. Rodríguez-García *et al.*, "Sentiment analysis on tweets about diabetes: An aspect-level approach," *Computational and Mathematical Methods in Medicine*, vol. 2017, pp. 1–9, 2017.

[62] S. Archondakis, M. Roma and K. Evropi, "Remote cytological diagnosis of salivary gland lesions by means of precaptured videos," *Journal of the American Society of Cytopathology*, PMID: 33707150, vol. 10, no. 4, pp. 435–443, 2021.

[63] T. Jensen, J. S. S. Secher and P. Kjaer, "Intra-and interobserver reproducibility of vertebral endplate signal (modic) changes in the lumbar spine: The nordic modic consensus group classification." *Acta Radiologica*, vol. 48, no. 7, pp. 748–754, 2007.