# An Approach to Sentiment Analysis on Gujarati Comments on Gujarati YouTube Video

Komal Vekariya, Mausam Dankhara, and Vipul Gamit

Babu Madhav Institute of Information Technology, Bardoli, India
14mscit092@gmail.com
Babu Madhav Institute of Information Technology, Bardoli, India
14mscit067@gmail.com
Babu Madhav Institute of Information Technology, Bardoli, India
vipul.gamit@utu.ac.in

## ABSTRACT

For the last several years, YouTube has being the sensation for social videos sharing, also high amount of people post their reviews on shared video. But manual analysis of review classification is very difficult to do and also became time consuming. Also sentiment analysis for English language has expanded very high but for Indian regional languages, the process is very slow. So this paper proposed a technical way for review classification on Gujarati comments posted by users on Gujarati videos. A general process that can be implemented in future for review classification on Gujarati review comment is brought with detailed description. This process will give positive, negative and neutral analysis for video review, so users can easily identify no. of positive, negative and neutral reviews on particular Guajarati video. Data used for Gujarati video review classification has been collected from Youtube.com. As a pre-processing technique tokenization, stop-word removal and stemming and for Feature Extraction POS (Part of Speech) tagging has been used. In this paper for Gujarati video feature analysis hybrid technique has proposed and graphical representation of feature analysis has shown by column chart.

**Keywords---** Sentiment Analysis, review Classification, Gujarati Comments, Machine Learning (ML).

## 1. INTRODUCTION

Sentiment includes feeling, emotion and way of thinking, attitude, and opinion towards some things, some people, some events, some action and many more. [4] Sentiment analysis is a multidisciplinary and multifaceted Artificial intelligence problem, its aim is to minimize the gap between human and computer, it is collection of human intelligence and electronic intelligence for mining the text and classifying user sentiments, likes, dislikes and wishes[1]. Sentiment analysis is also known as **Opinion Mining, Opinion Extraction, Sentiment Mining and Subjective Analysis** [7]**.**

In today's era where YouTube is the first choice of majority people to share a social video. On that shared video other users and viewers can also share their opinions, reviews using different languages that describes what the people think about overall video in actual. Look after review classification video owner can came to know that what the changes I have to do in my next upcoming video or user has liked and disliked these particular things on my video. Now manual analysis of comment is not possible, because there can be lac no. of comments available on a popular video. Also a lot of research has done on English review/comments.

For human being language is the best way to express their emotions towards the thing that has already happened or is happening or will happen, especially that regional language to which he/she is belonging. One of the regional language is Gujarati, spoken by over 55 million people worldwide (http://www.straitstimes.com). Also there hasn't much work done on Gujarati language, so this paper has proposed technique for sentiment analysis on Gujarati comments Using Gujarati YouTube video. At the last output will be categorized as 'positive', 'negative' and 'neutral' reviews about Gujarati video features and will be shown in the form of column graph. So that viewers and users can easily come to know the good, bad and neutral ratio about the video features.

## 1.1 Current Need

Sentiment is related to feelings and opinion of a particular person, that how they think about the thing. Now in a day's owner of the e-commerce and social sites allow users to express their feeling and opinion toward particular product, post, video, event, action and etc. It can be possible that there can be thousands of comments are available on particular shared post. So it will be almost difficult /impossible to do manual analysis on all of the comments. [11]

Using sentiment analysis comments can be classified into positive, negative and neutral. There are the methods that classify the comments and give total no. of positive, negative and neutral comments.

## 1.2 Sentiment Analysis is Classified into Three Main Levels [3]

### 1.2.1 Document Level Sentiment Analysis:

According to a particular topic or thing, a whole document is considered as either positive or negative or a neutral document, in document level sentiment analysis. They doesn't focus on sentence level analysis. Looking after overall analyze of document decision has made that in which category the whole document should be put.

### 1.2.2 Sentence Level Sentiment Analysis:

Each sentence is considered as either a positive, negative or neutral sentence. Sentiment analysis done on each sentence. [4]

### 1.2.3 Aspect Level Sentiment Analysis: [6]

It works better when combination of two sentences are there. The sentence has written to give review on a single feature/attribute, but it might be possible that in starting of that sentence he/she give positive review and then after or at the end of the comment it give negative review also on that attribute and vice versa.

Ex. 1. સફરજન નો લાલ કલર સારો છે, પણ સ્વાદ માં ગળ્યું નથી.

# 2. LITERATURE REVIEW

## 2.1 Sentiment Analysis Features

- Analyze Comments [8]

-Compare Review

-Product Improvement

-Better Marketing

## 2.2 Sentiment Analysis Stages [2]

There are multiple stages that describes overall process of sentiment analysis. Let's understand the process through a flowchart [2].

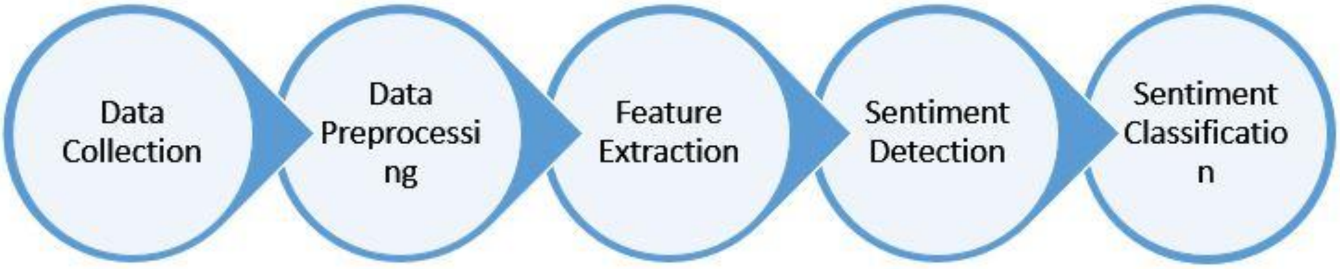


**Figure 1 Sentiment Analysis Stages**

Let's take a brief overview on each stage of sentiment analysis.

### 2.2.1 Data Collection:

To collect data for review analysis there are many resources like movie reviews, product reviews, traveler reviews, twitter twits, YouTube comments, Facebook posts comments etc. These collected data will be further proceed for pre-processing.

### 2.2.2 Data Pre-processing:

Data Pre-Processing is the technique to remove redundant or unwanted words from the given sentence, so it will make further process of feature extraction and sentiment detection very easier. There are many methods to perform data pre-store collection of sentences, it split the sentence into tokens, remove white space, unwanted symbols etc.

**Stop-word-Removal**- It remove extra words like હતો,હતી,છે,હશે,અને,તેનું,પેલું etc.

**Stemming-** Stemming algorithm reduces the work 'પેલીનું', 'પેલાનું' to the root word 'પેલું'.

### 2.2.3 Feature Extraction:

In this phase, remaining data proceed further after pre-processing phase. Feature extraction looks for the noun, adverb and adjective for further proceed. POS tagging is one of the mostly used method for feature extraction.

### 2.2.4 Sentiment Classification:

The basic principles of text classification based on machine learning methods are: the computer system automatically estimates the correlation between each text and various types of categories, then assign the text to one of the categories. Classification is done on the basis of three classes i.e. positive, negative and neutral. Various techniques are available like SVM, NB (Naïve Bayes), K-NN (K-Nearest Neighbor), NLP, MLP (Multi-layer perceptron) etc. [2]

### 2.2.5 Classified Output

Precision and Recall are usually used to measure the accuracy of a Sentiment Analysis methods. [5]

## 2.3. COMPARATIVE STUDY OF DIFFERENT SENTIMENT ANALYSIS TECHNIQUES

### 2.3.1 Machine Learning: [9]

In machine learning technique two documents are required: Test Dataset and Training Dataset. Dataset can be collection of more reviews or less reviews, as per your requirement .In processing like Stemming, Tokenization, Stop-word removal etc. After pre-processing data will be sent for feature extraction.

**Tokenization-** Tokenization is a text document in which it training dataset each sentence is already classified as either negative or positive. Where as in test dataset reviews are not classified as either positive or negative, you have to check it using different classification methods, so that you can get accuracy of different classifiers. There are mainly three classifiers which proves their accuracy greatest. Naive Bayes (NB), Maximum Entropy (ME) and Support Vector Machine (SVM).

**Dis-Advantage: [10]**

- Very low capable for new data.
- It is necessary that labeled data should be there, no matter either they are costly

### 2.3.2 Lexicon Based: [9]

Sentiment lexicon contains lists of words and expressions used to express people's subjective feelings and opinions. For example, start with positive and negative word lexicons, analyze the document for which sentiment need to find. Then if the document has more positive word lexicons, it is positive, otherwise it is negative. The lexicon based techniques to Sentiment analysis is unsupervised learning because it does not require prior training in order to classify the data.

There are three methods to construct a sentiment lexicon: manual construction, corpus-based methods and dictionary-based methods. The manual construction of sentiment lexicon is a difficult and time-consuming task.

-Use dictionary based approach (predefined list of words).

**Advantage:** wide no. of terms can used for better results.

**Dis-Advantage:** Only countable words are used. Doesn't provide the functionality to use infinite no. of words.

### 2.3.3 Hybrid Approach: [9]

Hybrid the combination of both the machine learning and the lexicon based approaches improve sentiment classification performance.

**TABLE-I**

| Methods | Sub-Methods | Advantage | Disadvantage |
|---|---|---|---|
| Machine Learning | Support Vector Machine[12] | -It gives high accuracy if the amount of data is huge. | -Dataset (size) requirement is very large. |

| | | | |
|---|---|---|---|
| | | -It works well even if data is not separate linearly.<br>-It is robust for sentiment classification.<br>-For multi-class classification this method is good to use. | -Memory consume for large dataset in manner of proper classification.<br>-As well require high speed in testing. |
| | Naive Bayes[12] | It is very simple to implement.<br>-Great computation efficiency.<br>-It provides accurate result for most of the classification. | -For good precision it requires large amount of data. |
| Lexicon Based | Dictionary Based Approach[13] | -We can easily find a good data set of sentiment words.<br>-We can expand manual dictionary using synonyms and antonyms. | -Low precise<br>Ex. Synonyms like Great=>excellent, admirable but also=>large, big, fat.<br>-It doesn't provide any domain specific list of words. |
| | Corpus Based Approach[13] | -We can expand it using grammar binding.<br>-It is able to obtain specific domain. | -It is hard to find a very large set of sentiment words.<br>- A list of word set/sentiment set is not complete.<br>-Requires large corpus to achieve good coverage. |
| Hybrid Approach | Both the combination of machine learning and lexicon based technique.[9] | -can get high accuracy from supervised powerful lexicon algorithm<br>-stability and reliability. | -Noisy reviews[10] |

Table-1: Comparative Study of Different Sentiment Analysis Techniques

## 3. PROPOSED MODEL

Diagram of introduced model is presented below.

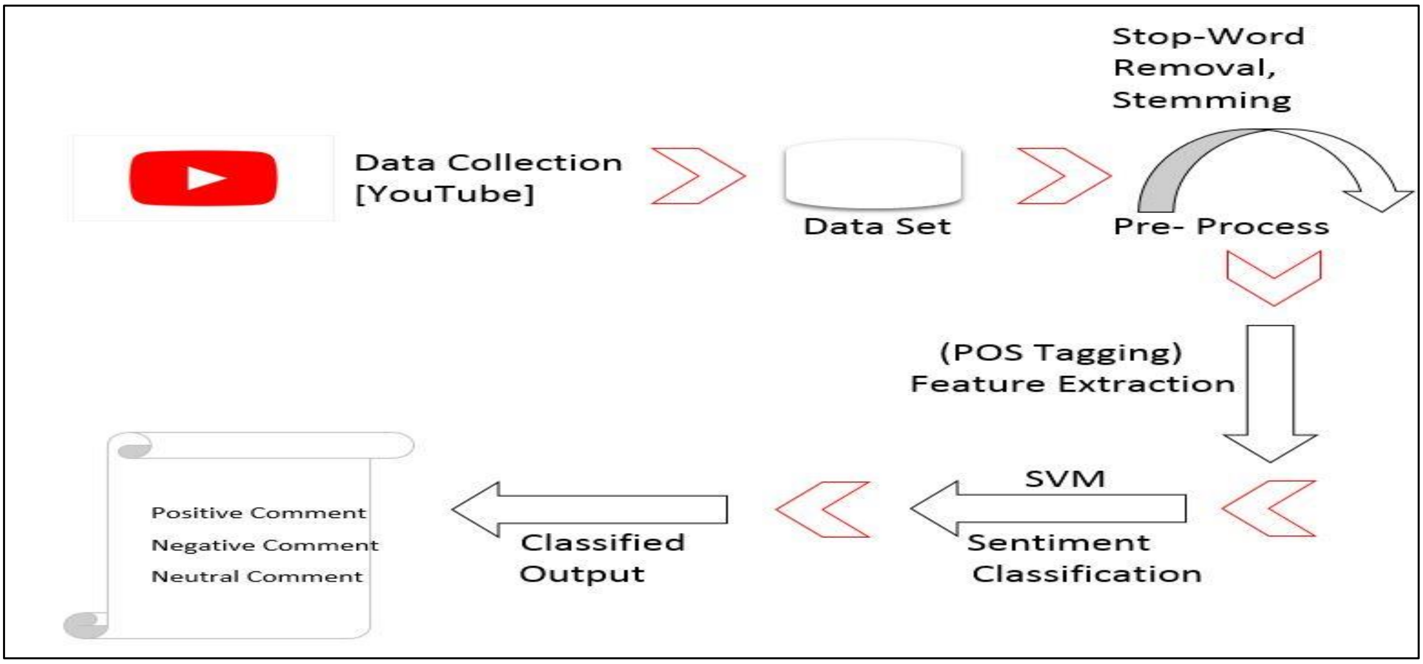


**Figure-2 Proposed Model [2]**

On below given comments we can perform all the steps given in proposed model.

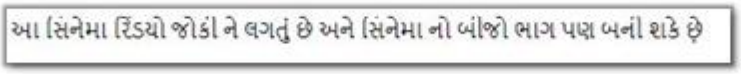
### 3.1 DATA PROCESSING

### 3.1.1 Data Collection:

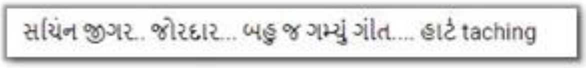
- Screen shot of Gujarati Comment:

The first step of sentiment analysis where user have to collect data from any knowable source.

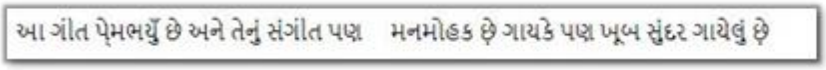
-Here we have collected data from Youtube.com, to do classification of Gujarati comments on Gujarati YouTube video.



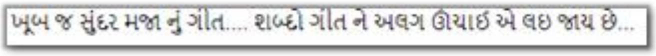
આ સિનેમા વિડિયો જોકો ને લગતું છે અને સિનેમા નો બીજો ભાગ પણ બની શકે છે



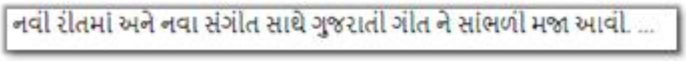
સચિન જીગર.. જોરદાર... બહુ જ ગમ્યું ગીત.... હાર્ટ taching



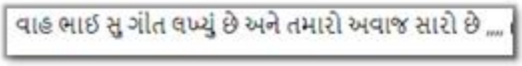
આ ગીત પ્રેમભર્યું છે અને તેનું સંગીત પણ મનમોહક છે ગાયકે પણ ખૂબ સુંદર ગાયેલું છે



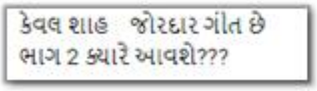
ખૂબ જ સુંદર મજા નું ગીત.... શબ્દો ગીત ને અલગ ઊંચાઈ એ લઇ જાય છે...



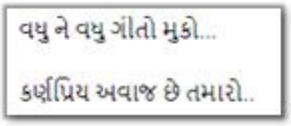
નવી રીતમાં અને નવા સંગીત સાથે ગુજરાતી ગીત ને સાંભળી મજા આવી. ...



વાહ ભાઈ સુ ગીત લખ્યું છે અને તમારો અવાજ સારો છે ....

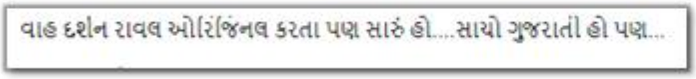


કેવલ શાહ જોરદાર ગીત છે
ભાગ 2 ક્યારે આવશે???

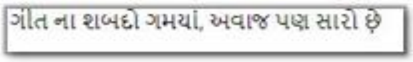


વધુ ને વધુ ગીતો મુકો...

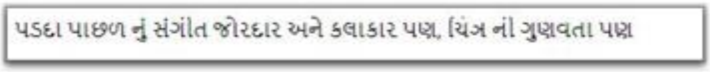
કર્ણપ્રિય અવાજ છે તમારો..



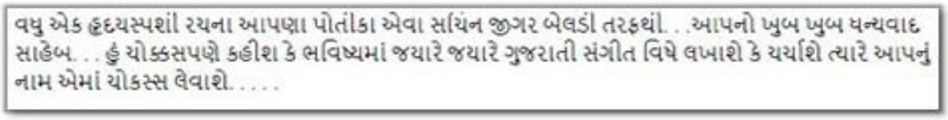
વાહ દર્શન રાવલ ઓરિજિનલ કરતા પણ સારું હો....સાચો ગુજરાતી હો પણ...



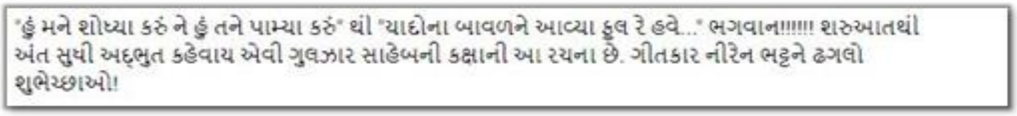
ગીત ના શબ્દો ગમયાં, અવાજ પણ સારો છે



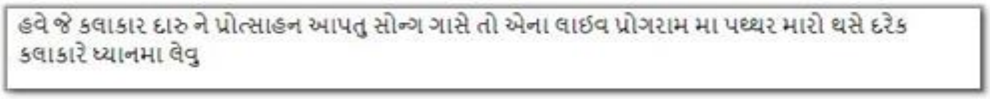
પડદા પાછળ નું સંગીત જોરદાર અને કલાકાર પણ, ચિત્ર ની ગુણવતા પણ



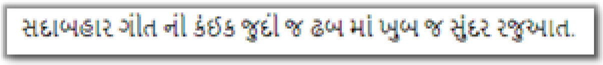
વધુ એક હૃદયસ્પર્શી રચના આપણા પોતીકા એવા સચિન જીગર બેલડી તરફથી. . .આપનો ખુબ ખુબ ધન્યવાદ સાહેબ. . . હું ચોક્કસપણે કહીશ કે ભવિષ્યમાં જયારે જયારે ગુજરાતી સંગીત વિષે લખાશે કે ચર્ચાશે ત્યારે આપનું નામ એમાં ચોકસ્સ લેવાશે. . . . .

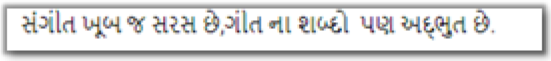


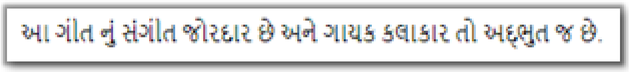
"હું મને શોધ્યા કરું ને હું તને પામ્યા કરું" થી "યાદોના બાવળને આવ્યા ફુલ રે હવે..." ભગવાન!!!!!! શરુઆતથી અંત સુધી અદ્દભુત કહેવાય એવી ગુલઝાર સાહેબની કક્ષાની આ રચના છે. ગીતકાર નીરેન ભટ્ટને ઢગલો શુભેચ્છાઓ!

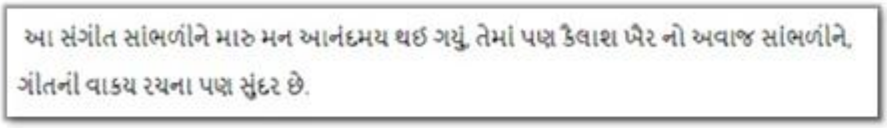


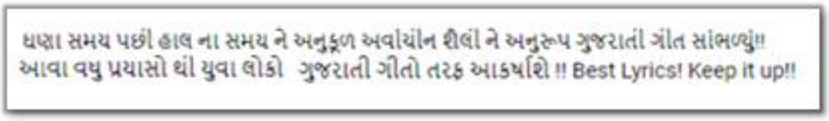
હવે જે કલાકાર દારુ ને પ્રોત્સાહન આપતુ સોન્ગ ગાસે તો એના લાઈવ પ્રોગરામ મા પથ્થર મારો થસે દરેક કલાકારે ધ્યાનમા લેવુ


સદાબહાર ગીત નો કંઈક જુદો જ ઢબ માં ખુબ જ સુંદર રજુઆત.


સંગીત ખૂબ જ સરસ છે,ગીત ના શબ્દો  પણ અદ્દ્ભુત છે.


આ ગીત નું સંગીત જોરદાર છે અને ગાયક કલાકાર તો અદ્દ્ભુત જ છે.


આ સંગીત સાંભળીને મારુ મન આનંદમય થઇ ગયું, તેમાં પણ કૈલાશ ખેર નો અવાજ સાંભળીને,
ગીતની વાકય રચના પણ સુંદર છે.


ઘણા સમય પછી હાલ ના સમય ને અનુકૂળ અર્વાચીન શૈલી ને અનુરૂપ ગુજરાતી ગીત સાંભળ્યું!!
આવા વધુ પ્રયાસો થી યુવા લોકો  ગુજરાતી ગીતો તરફ આકર્ષશે !! Best Lyrics! Keep it up!!

### 3.1.2 Data Pre-processing:

In this step we will remove extra and unwanted words through different techniques for better classification.

#### 3.1.2.1 Tokenization:

Before tokenization unnecessary symbols are removed trough regular expression and we will look only for Gujarati characters.

**Ex.1** **[-** વધુ એક હદયસ્પર્શી રચના આપના પોતીકા એવા સચિન જીગર બેલડી તરફથી

- આપનો ખુબ ખુબ ધન્યવાદ સાહેબ

- હું ચોકકસપણે કહીશ કે ભવિષ્યમાં જ્યારે જ્યારે ગુજરાતી સંગીત વિષે લખાશે કે ચર્ચાશે ત્યારે આપનું નામ  એમાં ચોક્કસ લેવાશે.  **]**

**Ex.2** **[-** ગુજરાત ની સંસ્કૃતિ ભસ્મ કરી નાખી આવા સોંગ રિલીસ કરીને

- થુ**]**

-Here you can see the difference between above comment and screen shot comment , we have consider[…] as an end of line, and after word will consider as a new line.

**Ex.3** ગીત ના શબ્દો ગમયા આવાજ પણ સારો છે.

**Ex.4** પડદા પાછળનું સંગીત જોરદાર અને કલાકાર પણ ચિત્ર ની ગુણવત્તા પણ

-Here it will remove [,] and make a line as a single line and will also remove [.] from at the end of the line

-Now it will convert each sentence into tokens using **Ex.1**:

[- "વધુ" "એક" "હદયસ્પર્શી" "રચના" "આપના" "પોતીકા" "એવા" "સચિન" "જીગર" "બેલડી" "તરફથી"

- "આપનો" "ખુબ" "ખુબ" "ધન્યવાદ" "સાહેબ"

- "હું" "ચોક્કસપણે" "કહીશ" "કે" "ભવિષ્યમાં" "જ્યારે" "જ્યારે" "ગુજરાતી" "સંગીત" "વિષે" "લખાશે" "કે" "ચર્ચાશે" "ત્યારે" "આપનું" "નામ" "એમાં" **"**ચોક્કસ**"** **"**લેવાશે**"]**

#### 3.1.2.2 Stop Word Removal:

In this technique it will remove extra and unwanted words like:

**Ex.1** [- "વધુ" "એક" "હદયસ્પર્શી" "રચના" "પોતીકા" "સચિન" "જીગર" "બેલડી"

- “ખુબ” “ખુબ” “ધન્યવાદ” “સાહેબ”
- “ચોક્કસપણે” “ભવિષ્યમાં” “ગુજરાતી” “સંગીત” “લખાશે” “ચર્ચાશે” “નામ” “ચોક્કસ” “લેવાશે”]

**Ex.2** [“પડદા” “પાછળનું” “સંગીત” “જોરદાર” “કલાકાર” “ચિત્ર” “ની” “ગુણવત્તા”]

### 3.1.2.3 Stemming:

In this technique it will remove extra characters like પાસેની માંથી ની નીકળી જશે, ત્યાંથી માંથી થી નીકળી જશે,પેલાનો માંથી નો અને એના જેવા જ બીજા શબ્દો જેવા કે ના, ની, ને, નું નીકળી જશે.

**Ex.1** [- “વધુ” “એક” “હદયસ્પર્શ” “રચના” “પોતીકા” “સચિન” “જીગર” “બેલડી”

- “ખુબ” “ખુબ” “ધન્યવાદ” “સાહેબ”
- “ચોક્કસ” “ભવિષ્ય” “ગુજરાતી” “સંગીત” “લખ” “ચર્ચા” “નામ” “ચોક્કસ” “લેવ”]

**Ex.2** **[-** “પડદા” “પાછળ” “સંગીત” “જોરદાર” “કલાકાર” “ચિત્ર” “ગુણવત્તા”**]**

### 3.1.3 Feature Extraction:

**Ex.**

| | |
|---|---|
| **Noun** | સચિન, જીગર, સાહેબ, ગુજરાતી, રચના, સંગીત, નામ, પડદા, કલાકાર, ચિત્ર, બેલડી |
| **Adverb** | એક, ખુબ, વધુ, ચર્ચા |
| **Adjective** | હદયસ્પર્શ, ધન્યવાદ, પોતીકા, ચોક્કસ, પાછળ, જોરદાર, ગુણવત્તા, ભવિષ્ય |

### 3.1.4. Sentiment Classification

In given approach they have used SVM classifier to classify the comment.

Now here we can use dictionary based approach, with SVM.

-SVM will first classify into specific three categories positive, negative and neutral.

-Now using dictionary based approach we can find all the Gujarati video regarding features we have put in dictionary We have to find each feature in all the thrice categories.

-Ex. There is a Gujarati song video and one of its feature named as ગીત.

Now we have to find all the comments related to feature ગીત in all the thrice categories.

-After that we will count total no. of positive, negative and neutral comments regard to the features ગીત and the graph will create like this:

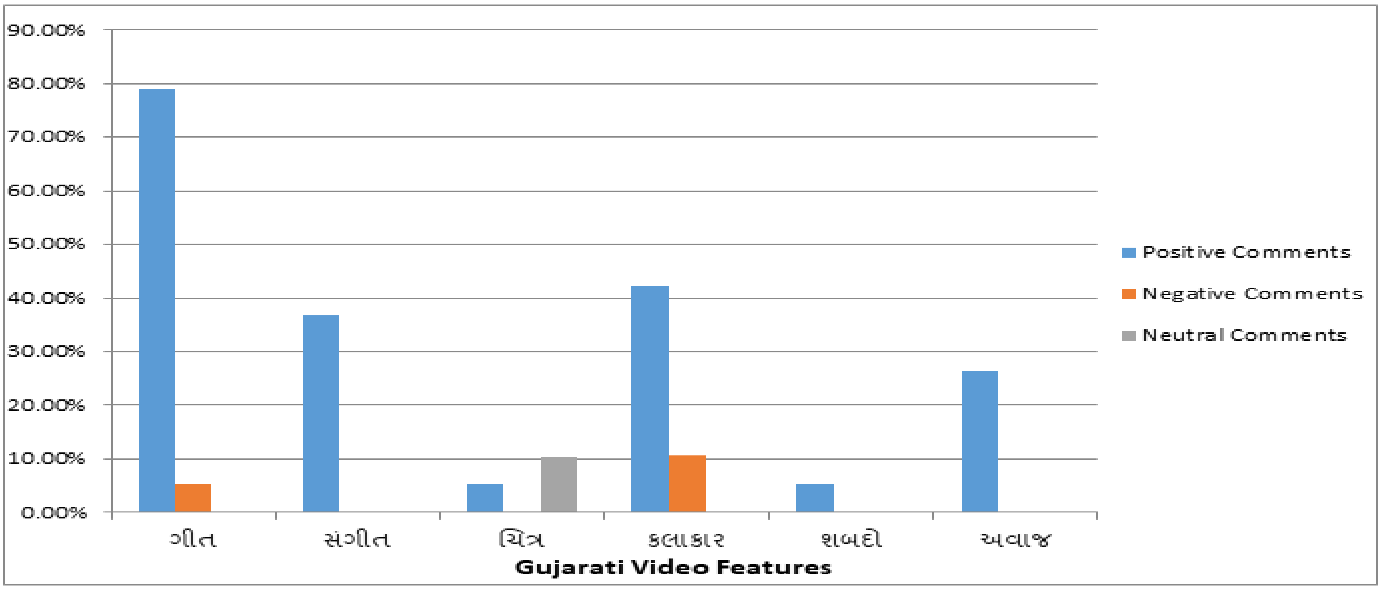


**Figure-3 Analysis of Gujarati Video Comments**

-As you can see we have listed some of the features of Gujarati video song and we have find these listed features from the comments given in section[v] of literature review.

-The graph shows total no. of positive, negative and neutral comments on each feature out of all the comments.

## 4. CONCLUSIONS

After analysis and comparison of all the thrice methods of sentiment analysis we came to the conclusion that in future we will use Hybrid Approach for sentiment classification. Because we will perform analysis on Gujarat video comment, we require predefined word of list for finding all the features of video and sentiment words, also we will use a dataset. So according to our need for classification purpose machine learning technique will be suited for us and to compare predefined list of sentiment word with dataset lexicon method will be preferable to us. So the both combination of machine learning and lexicon known as Hybrid approach will be used.

## REFERENCES

[1] Amandeep Kaur, Vishal Gupta “A Survey on Sentiment Analysis and Opinion Mining Techniques”, VOL. 5, NO. 4, NOVEMBER 2013.

[2] Dr. Vipul M. Vekariya, Vrunda C. Joshi “An Approach to Sentiment Analysis on Gujarati Tweets”, ISSN 0973-6107 Volume 10, Number 5 (2017) pp. 1487-1493.

[3] Neha Nehra “A SURVEY ON SENTIMENT ANALYSIS OF MOVIE REVIEWS”, 2014 IJIRT | Volume 1 Issue 7 | ISSN: 2394-6002.

[4] Muhammad Zubair Asghar1, Aurangzeb Khan2, Shakeel Ahmad1, Fazal Masud Kundi1 “A Review of Feature Extraction in Sentiment Analysis”, *J. Basic. Appl. Sci. Res.*, 4(3)181-186, 2014**,** ISSN 2090-4304.

[5] Muhammad Zubair Asghar1, Shakeel Ahmad2, Afsana Marwat1, Fazal Masud Kundi1,” Sentiment Analysis on YouTube: A Brief Survey”.

[6] 1Thellaamudhan C, 2Suresh R, 3Raghavi P “A Comprehensive Survey on Aspect Based Sentiment Analysis”, Volume 6, Issue 4, April 2016 , ISSN: 2277 128X

[7] S. Kasthuri1, Dr. L. Jayasimman2, Dr. A. Nisha Jebaseeli3, “An Opinion Mining and Sentiment Analysis Techniques: A Survey”, e-ISSN: 2395 -0056, Volume: 03 Issue: 02 | Feb-2016, p-ISSN: 2395-0072

[8] Devare Jayvant1 ,Punde Sunita2,Sahane Sujata3,Shendage Sonali4 and Kanase Rajesh5 “Product Review By Sentiment Analysis” Volume 3 Issue 5 may, 2014 Page No. 6202-6205, ISSN:2319-7242

[9] MR. S. M. VOHRA, 2 PROF. J. B. TERAIYA,”A COMPARATIVE STUDY OF SENTIMENT ANALYSIS TECHNIQUES” ISSN: 0975 – 6760| NOV 12 TO OCT 13 | VOLUME – 02, ISSUE – 02.

[10] Alessia D’Andrea, Fernando Ferri, Patrizia Grifoni, Tiziana Guzzo, “Approaches, Tools and Applications for Sentiment Analysis Implementation”, International Journal of Computer Applications (0975 – 8887) Volume 125 – No.3, September 2015.

[11] https://www.slideshare.net/makrandp/introduction-27376010

[12] Mohini Chaudhari, Sharvari Govilkar, “A SURVEY OF MACHINE LEARNING TECHNIQUES FOR SENTIMENT CLASSIFICATION”, International Journal on Computational Sciences & Applications (IJCSA) Vol.5, No.3, June 2015 DOI: 10.5121/.

[13] Prof. Ronen Feldman, “SENTIMENT ANALYSIS TUTORIAL”, Digital Trowel, Empire State Building.