International Conference on Advances in Management & Technology (ICAMT- March 2025)

RESEARCH ARTICLE OPEN ACCESS

Comparative Study of Machine Learning Algorithms for Sentiment Analysis

Miss.Ritika.M.Ganvir[1]
[1]Shri Shivaji Education Society Amravati's Science College, Nagpur

Miss.Rasika.D.Bagal[2]
[2]Shri Shivaji Education Society Amravati's Science College, Nagpur

Miss.AsfiyaShaikh[3]
[3]Shri Shivaji Education Society Amravati's Science College, Nagpur

Miss.AditiCharde[4]
[4]Shri Shivaji Education Society Amravati's Science College, Nagpur

Miss.SonaliThakur[5]
[5]Shri Shivaji Education SocietyAmravati's Science College, Nagpur

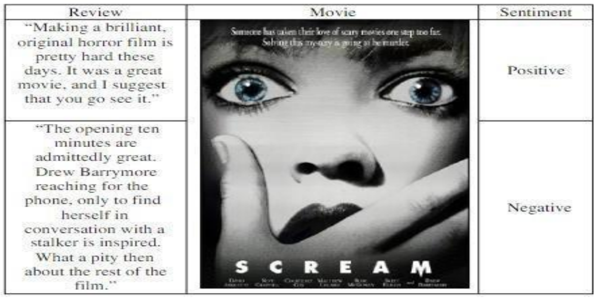
ABSTRACT

The most studied topic is sentiment analysis. region these days. Before it can be utilised, a lot of information on the internet must be examined. Deciphering these data has been the focus of numerous researchers. The goal of sentiment analysis techniques is to reveal hidden feelings, ideas, or subjectivity inside a document. Machine learning techniques are used to analyse sentiment. In order to give context, this study looks at recently published research sentiment analysis, which is categorized based on its information extraction tasks. The issues encountered and potential difficulties with this research topic are also looked at and explored. The purpose of the Sentiment Analysis tool is to analyse a set of expressions according to their characteristics and quality. Because there are so many opinions, sentiment analysis is also known as opinion mining.

Introduction

Since everyone has a different perspective on everything, sentiment analysis is a popular topic. All people in the modern world use online platforms and frequently offer their thoughts, ideas, and comments. Numerous machine learning approaches can effectively mine this type of information.

"Sentiment analysis, additionally known as opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities like products, services, organizations, individuals, issues, events, topics, and their attributes". Sentiment analysis is an example of analysing texts and natural language Processing. looking for & extract arbitary information to the original origin via Processing and computational linguistics Analysis of sentiment is frequently applied to ratings and networking sites for a range of purposes, from customer support to marketing. A primary aspect of sentiment analysis is to determine if the viewpoint articulated in a statement, document, or other item exhibiting a particular characteristic is positive, negative, or neutral. This is known as categorising a text's orientation at the document level, phrase, or feature level. Sentiment analysis additionally categorises the statement beforehand based on emotional states like "happy,""sad,"and"angry"[1].



Review	Movie	Sentiment
"Making a brilliant, original horror film is pretty hard these days. It was a great movie, and I suggest that you go see it."		Positive
"The opening ten minutes are admittedly great. Drew Barrymore reaching for the phone, only to find herself in conversation with a stalker is inspired. What a pity then about the rest of the film."		Negative

Someone has taken their love of scary movies one step too far. Solving this mystery is going to be murder.
SCREAM

Fig : Glimpse of the dataset positive and negative reviews on the same movie.

Literature Survey

Numerous research publications are examined under the Literature Survey, which gives a brief overview of the research topic to the investigator. This section covers related work based on author reviews based on newly developed technologies and new trends that are connected to each other.

Agarwal (2015): It was discovered that identifying excellent features is a difficult challenge for improved outcomes while implementing machine learning techniques. The After the introduction of the "Semantic Parser" idea, concepts were treated as features. The The minimum redundancy and maximum relevance feature selection methodology (mRMR) . For their classification work, they employed various feature sets, such as dependency parse trees, bit agged, bigrams, and unigrams, in addition to their suggested plan in order to compare the results. [3]

Hassan Khan (2016): The method consists of meticulous data pre-processing as well as supervised machine learning. To ensure that machine learning is not restricted to a single domain, they gathered tagged datasets from other fields. They use various training sets to teach SVM classifiers, each of which teaches SVM a distinct collection of features. 1) Information gain(IG) when features are present, and 2) The frequency of features 3) cosine similarity with the existence of features, and 4) frequency of features. They discovered that feature frequency is inferior to feature presence. [4]

Mohey et al (2016): It mostly highlights the different classifiers' SA troubles. Overheads for NLP, feature extraction, and negation handling provide difficulties. The accuracy rate provides the basis for the second comparison. Here, a comparison is conducted using the most recent methods for sentiment analysis. [5]

Yang et al (2018): Accuracy, precision, recall, and F1 Score criteria are applied to evaluate the performance . SA analysis methods are in English language. [6]

Design and Implementation:

Work Flow Diagram

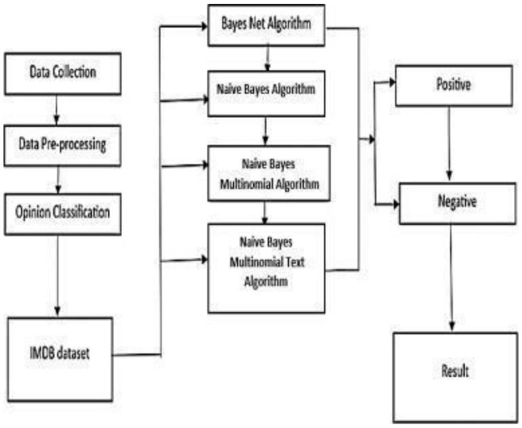


Fig: Work Flow Diagram.

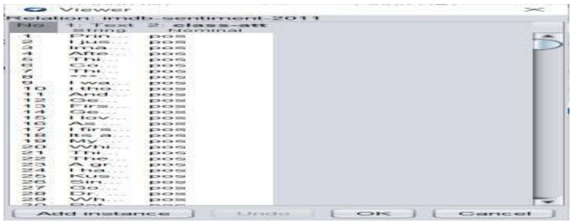
Methodology:

Supervised learning: It is used as a machine learning algorithm, its defined to usage of labelled datasets to train algorithms for reliable information classification or outcome prediction.

Unsupervised learning It is also called as "unsupervised machine learning", it analyses and groups unlabelled at a sets using machine learning algorithms. The detailed steps of our project evaluation are as follows.

Dataset and attribute selection

We have gathered a dataset called "IMDB" that includes the students' results from the previous semester. There are two attributes and 50,000 instances in the dataset. Additionally, there are several missing values. The data file must be in one of two formats: "CSV" or "ARFF" (Attribute Relation File Format). This is a sample of our dataset, which is 61.8 MB in size and in the "ARFF" format.



Relation: imdb-sentiment-2011

No.	1: Text String	2: class-att Nominal
1	Prin...	pos
2	I jus...	pos
3	Ima...	pos
4	Afte...	pos
5	Thi...	pos
6	Co...	pos
7	Thi...	pos
8	[illegible]	pos
9	I wa...	pos
10	I tho...	pos
11	And...	pos
12	Ge...	pos
13	Firs...	pos
14	Ge...	pos
15	I lov...	pos
16	As...	pos
17	I firs...	pos
18	Its a...	pos
19	My...	pos
20	Whi...	pos
21	Thi...	pos
22	The...	pos
23	A gr...	pos
24	I ha...	pos
25	Kus...	pos
26	Sin...	pos
27	Go...	pos
28	Dr...	pos
29	Wh...	pos

Fig: IMDB Dataset.

Filters

The pre-process section enables the definition of filters that alter the data in different ways. The necessary filters are set up using the filter box. The two primary types of filters are supervised and unsupervised. In this case, unsupervised category filters will be used. If the dataset contains any information that string attribute convert into set of numeric attributes representing word by using **'String to Word**

Vector' filter under attribute section of unsupervised filters.

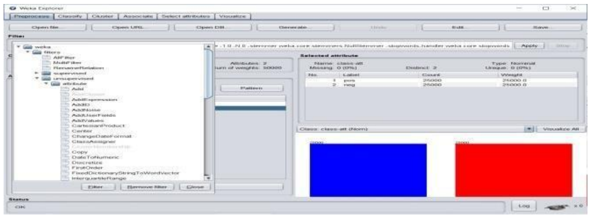
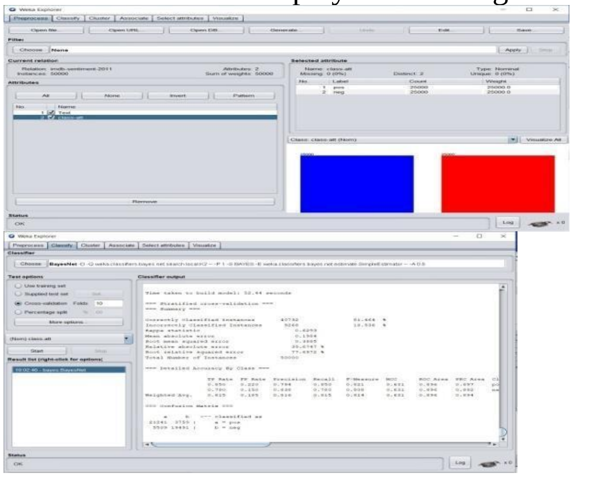


Fig: Filtering of Dataset.

Pre-processing

Data Pre-processing is the initial stage of this project's assessment. We will use the WEKA Explorer interface for our project. Here, the local computer is used to choose the source data file. We can refine the data by choosing several options, referred to as "Data Cleaning," after loading it into Explorer. We can also choose or delete properties based on our needs. The pre- processed version of our dataset is as follows. The relation name, number of attributes, and number of records are displayed in detail on the left-hand side of the screen above. Details of attribute values are displayed on the right side.



```
Time taken to build model: 52.44 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       40732               81.464  %
Incorrectly Classified Instances      9268               18.536  %
Kappa statistic                          0.6293
Mean absolute error                      0.1984
Root mean squared error                  0.3885
Relative absolute error                 39.6747 %
Root relative squared error             77.6872 %
Total Number of Instances            50000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cl
                 0.850    0.220    0.794      0.850    0.821      0.631    0.896     0.897     po
                 0.780    0.150    0.839      0.780    0.808      0.631    0.896     0.892     ne
Weighted Avg.    0.815    0.185    0.816      0.815    0.814      0.631    0.896     0.894

=== Confusion Matrix ===

     a     b   <-- classified as
 21241  3759 |     a = pos
  5509 19491 |     b = neg
```

Fig: Preprocessing.

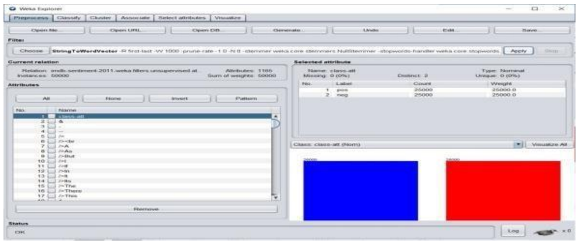


Fig: Total No. of Attributes and instances.

Here the total number of attributes is 1165 and instances 50000. This dataset is evaluating negative values is 25000 and positive values is 25000. Visualize all to shows the mini graph to all attributes.

Fig:Mini Graph of All Attributes.

Classification

For our prediction purpose we have to choose a classifier. Select the attribute to use as the class [(nom)@@class@@].We will select a standard classifier named as bayes net, naive bayes, for classification.

Baye s Net

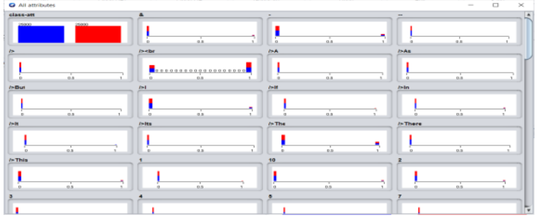
Every node in a Bayesian network indicates a variable that is random, and every limit shows the associated random variables' conditional probability. An uncertain domain's knowledge can be expressed using a probabilistic graphical model termed (BN). Bayes networks and belief networks are other names for BNs.

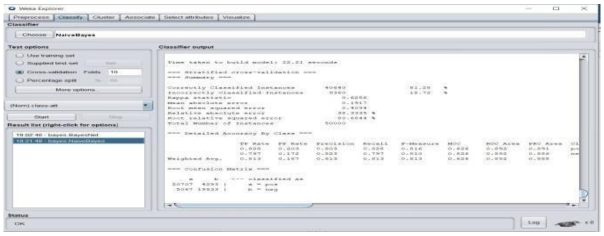
Fig: Output of Bayes Net Algorithm.

Fig.Output of Bayes Net Algorithm.

Naïve Bayes

Based on the Bayes theorem, Classification issues are resolved through the Naive Bayes algorithm, a supervised learning technique. Text grouping using a high-dimensional training dataset is its main use case. The Naive Bayes Classifier, one of the most simple and efficient categorisation algorithms, facilitates the quick development of models for machine learning with predictive powers. As a classifier based on probability, it bases its forecasts regarding the basic principles of the item likelihood. Sentiment analysis, article classification, and spam filtering are a few well- known applications of the Naïve Bayes algorithm.



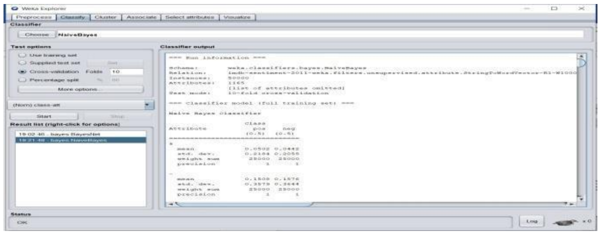


```
Time taken to build model: 22.21 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       40640               81.28   %
Incorrectly Classified Instances      9360               18.72   %
Kappa statistic                          0.6256
Mean absolute error                      0.1917
Root mean squared error                  0.4034
Relative absolute error                 38.3335 %
Root relative squared error             80.6844 %
Total Number of Instances            50000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cl
                 0.828    0.203    0.803      0.828    0.816      0.626    0.892     0.891     po
                 0.797    0.172    0.823      0.797    0.810      0.626    0.892     0.886     ne
Weighted Avg.    0.813    0.187    0.813      0.813    0.813      0.626    0.892     0.888

=== Confusion Matrix ===

     a     b   <-- classified as
 20707  4293 |     a = pos
  5067 19933 |     b = neg
```

Fig: Output of Naïve Bayes Algorithm.

Fig: Output of Naïve Bayes Algorithm.

Performance and results



```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     imdb-sentiment-2011-weka.filters.unsupervised.attribute.StringToWordVector-R1-W1000
Instances:    50000
Attributes:   1165
              [list of attributes omitted]
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

                    Class
Attribute             pos     neg
                    (0.5)   (0.5)
==================================
&
  mean             0.0502  0.0442
  std. dev.        0.2184  0.2055
  weight sum        25000   25000
  precision             1       1

-
  mean             0.1508  0.1576
  std. dev.        0.3579  0.3644
  weight sum        25000   25000
  precision             1       1
```

Bayes net Output:

```
===Run information==

Scheme:      weka.classifiers.bayes.BayesNet -D-Qweka.classifiers.bayes.net.search.local.K2---P 1
-S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
Relation:         imdb-sentiment-2011-weka.filters.unsupervised.attribute.StringToWordVector-R1-
W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-
handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer-delimiters"
\r\n\t.,;:\'\"()?!"
Instances:    50000
Attributes:1165
            [list of attributes omitted]
Test mode:   10-foldcross-validation
```

```
===DetailedAccuracyByClass===

TP RateFPRatePrecisionRecallF-MeasureMCCROC AreaPRC AreaClass 0.828
           0.203     0.803    0.828   0.816         0.626   0.892        0.891      pos
0.797      0.172     0.823    0.797   0.810         0.626   0.892        0.886      neg
WeightedAvg.   0.813   0.187   0.813    0.813   0.813    0.626   0.892    0.888

===ConfusionMatrix===

   a    b<--classifiedas
 207074293 |      a = pos
  506719933|      b =neg
```

Naïve bayes Output:

```
===DetailedAccuracyByClass===

TP RateFPRatePrecisionRecallF-MeasureMCCROCAreaPRCAreaClass 0.850
            0.220     0.794    0.850   0.821         0.631   0.896         0.897      pos
0.780       0.150     0.838     0.780   0.808         0.631   0.896         0.892      neg
WeightedAvg.   0.815   0.185   0.816    0.815   0.814    0.631   0.896    0.894

===ConfusionMatrix ===++++++

   a    b<--classifiedas
 212413759 |      a = pos
  550919491|      b =neg
===Run information===

Scheme:      weka.classifiers.bayes.Naive Bayes
Relation:         imdb-sentiment-2011-weka.filters.unsupervised.attribute.StringToWordVector-R1-
W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-
handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer-delimiters"
\r\n\t.,;:\'\"()?!"
Instances:    50000
Attributes:1165
            [list of attributes omitted]
Test mode:   10-foldcross-validation
```

```
===Classifier model(full trainingset)===

Naive Bayes Classifier

                 Class
Attribute              pos    neg
                 (0.5)(0.5)
====================================================
&
  mean             0.05020.0442
  std.dev.         0.21840.2055
  weightsum              2500025000
  precision              1      1

-
  mean             0.15080.1576
  std.dev.         0.35790.3644
  weightsum              2500025000
  precision              1      1

mean            0.0292 0.0327
std.dev.         0.16840.1779
weightsum           2500025000
precision             1      1
```

Margin Curve Graph

Right side to click for option result list to click on right button and to shows the option visualized margin curve. All graph shows in negative and positive values. The blue color shows in negative value curve and orange colour show in positive value curve .

```
   ===Classifiermodel(fulltrainingset) ===

   Bayes Network Classifier
   not using ADTree
   #attributes=1165#classindex=0
   Network structure(nodes followed by parents)
   class-att(2):
   &(1):class-att
   -(1):class-att
   --(1):class-att
   /≥(1):class-att
   /><br(2):class-att
   />A(1):class-att
   .
===Stratifiedcross-validation===
===Summary===

CorrectlyClassifiedInstances       40732          81.464%
IncorrectlyClassifiedInstances      9268          18.536%
Kappa statistic                  0.6293
Meanabsoluteerror                   0.1984
Root meansquarederror                 0.3885
Relativeabsoluteerror             39.6747%
Rootrelativesquarederror            77.6972%
TotalNumberofInstances             50000
```

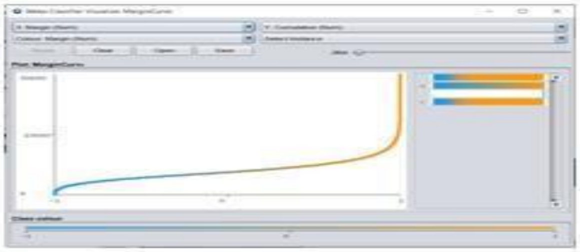


Fig: Bayes Net Graph

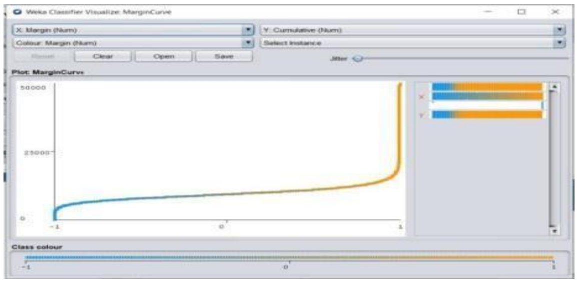


Fig.Naive Bayes Graph

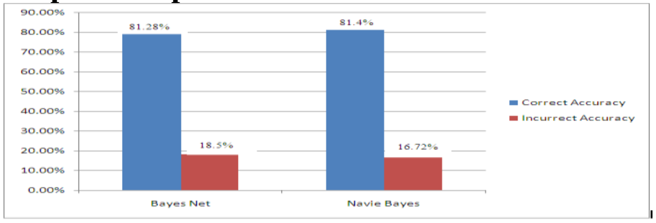
Result

The below tables hows the best performance is Naïve Bayes, because correctly accuracy is highest as compare to other algorithm and time taken consuming.

Sr. No	Algorithm	Accuracyof correctly classified instance	Instance	Accuracyof incorrectly classified instance	Instance	Total Instance	Time Taken
1	BayesNet	81.4%	40732	18.5%	9268	50000	52.44sec
2	NaiveBayes	81.28%	40640	16.72%	5360	50000	22.21sec

Sr. No	Algorithm	Precision	Recall	F-measure	Positive	Negative
1	BayesNet	0.816	0.815	0.814	21241 3759	5509 19491
2	NaiveBayes	0.813	0.813	0.813	20707 4293	5067 19933

GraphicalRepresentation



Conclusion

For performing Sentiment Analysis, on the IMDB dataset, we have applied total four variations of Naïve bayes algorithm i.e. Bayes net, Naive Bayes,. Out of which, better performance is shown by Naive Bayes algorithm as compare to other algorithms. Naive Bayes algorithm takes less time to give the best accuracy. As shown in table and graph the Naive Bayes algorithm is giving the best performance as compare to other algorithms, because accuracy of correctly classified instance is high as compare to another algorithm. Accuracy of correctly classified instance of Naive Bayes Text is 81.28% and it takes 22.21 seconds. One of the popular areas of web mining is opinion mining. Opinion mining offers numerous advantages for both customers and businesses.

References

1. Agnihotri, R., R. Dingus, M.Y. Hu and M.T. Krush, 2015.Social media: Influencing customer satisfaction inB2B sales. Industrial Market. Manage. DOI: 10.1016/j.indmarman.2015.09.003
2. Online Consumer Behaviour in Social Media Post Types: A Data Mining Approach Dimitrios Gkikas University of Patras The odoros The odoridis University of Salford Prokopis The odoridis University of

Patras Androniki Kavoura Prof. University of West Attica.

3. Review of Data Mining with Weka Tool Kulwinder Kaur1*, Shivani Dhiman2 1Department

of Computer Science Engineering, Indus International University Una, India2 Department of Computer Applications, Indus International University Una, India.

4. Opinion-Mining Methodology for Social Media Analytics Yoosin Kim1 , Seung Ryul Jeong21 College of Business, University of Texas at Arlington, Texas, US E-mail: yoosink@uta.edu
2 Graduate School of Business IT, Kookmin University, Seoul, Korea E- mail: srjeong@kookmin.ac.kr Corresponding Author: Seung Ryul Jeong Received October 23, 2014; revisedNovember23,2014;acceptedDecember 2,2014;publishedJanuary31,2014.

5. Opinion Mining on Social Media Data: Sentiment Analysis of User Preferences Vasile Daniel Păvăloaia 1,*,Elena- Mădălina Teodor2, Faculty of Business Information Technology and Statistics, Department of Accounting, Doina Fotache 1 and Magdalen Danile Economics and Business Administration, Alexandru Ioan Cuza University of Iasi, 700506 Ia¸si, Romania 2 Web Department, Falcon Trading Company, 700521 Ia¸si, Romania 3 Department of Management, Marketing and Business Administration, Faculty of Economics and Business Administration, Alexandru Ioan Cuza University of Iasi, Faculty of Economics and Business Administration, 700506 Iaşi, Romania * Email address: danpav@uaic.ro Received on July 23, 2019, accepted on August 14, 2019, and published on August 17.

6. The International Journal of Recent Technology and Engineering (IJRTE), Volume 8, Issue 4, November 2019, 9727, ISSN: 2277-3878 D9211118419/2019©BEIESP DOI: 10.35940/ijrte.D9211.118419 is the reference number. Blue Eyes Intelligence Engineering & Sciences Publication is the publisher. Machine learning-based general-purpose opinion mining system for social media sites Dhanraj Verma and Api Urmita Mehta.