A COMPUTER SCIENCE PORTAL WITH SENTIMENT ANAIYSIS

Ms. M.S.Namose[1], Ms.Srushti Mate[2], Mr.Kiran Mishra[3], Mr.Ashutosh Jagtap[4],
Mr.Shubham Pathak[5] [1]Professor, [2,3,4&5] Students
Department of computer science
JSPM Narhe Technical Campus Pune, Maharashtra. India

ABSTRACT

Our proposed system is computer science portal with sentiment analysis we are aiming to design an efficient portal where one can write article on various topic related to Computer Science field. It also consist of a forum where one can raise doubt. In Proposed System we are going to apply sentiment analysis on article's which will give learner an idea which article is better. The article are going to be sorted based on positive sentiment. Organizations use surveys, opinion polls, and social media as a mechanism to obtain feedback on their products and services. Sentiment analysis or opinion mining is the computational study of opinions, sentiments, and emotions expressed in text

Keywords: machine learning, sentiment analysis, social media

I. INTRODUCTION

Sentiment analysis is the process of using natural language processing, text analysis, and statistics to analyze customer sentiment. A basic task in sentiment analysis is classifying the *polarity* of a given text at the document, sentence, or feature/aspect level-whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral Sentiment analysis is done using algorithms that use text analysis and natural language processing to classify words as either positive, negative, or neutral.This allows companies to gain an overview of how their customers feel about the brand. Social networking portals have been widely used for expressing opinions in the public domain

Sentiment Analysis is one of the key emerging technologies in the effort to help people navigate the huge amount of user generated content available online. It is true that via these media citizens can express their desires, problems , emotions and feelings and the experts can make use of it by properly extracting and analyzing it. But the extraction and analysis of huge unstructured internet content is beyond the human power and time. The content is mostly written in natural language. This situation necessitate an automatic natural language processing tools that extract and analyze the people sentiments from this unstructured texts.Numerous researches are under going this direction. This research domain is called Opinion mining and sentiment analysis

Sentiment analysis is the process of using natural language processing, text analysis, and statistics to analyze customer sentiment. The best businesses understand the sentiment of their customers—what people are saying, how they're saying it, and what they mean. Customer sentiment can be found in tweets, comments, reviews, or other places where people mention your brand. Sentiment Analysis is the domain of understanding these emotions with software, and it's a must-understand for developers and business leaders in a modern workplace. As with many other fields, advances in deep learning have brought sentiment analysis into the foreground of cutting-edge algorithms. Today we use natural language processing, statistics, and text analysis to extract, and identify the sentiment of words into positive, negative, or neutral categories.

Sentiment analysis runs into a similar set of problems as emotion recognition does – before deciding what the sentiment of a given sentence is, we need to figure out what "sentiment" is in the first place. Is it categorical, and sentiment can be split into clear buckets like happy, sad, angry, or bored? Or is it dimensional, and sentiment needs to be evaluated on some sort of bi-directional spectrum?

In addition to the definition problem, there are multiple layers of meaning in any human-generated sentence. People express opinions in complex ways; rhetorical devices like sarcasm, irony, and implied meaning can mislead sentiment analysis. The only way to really understand these devices are through context: knowing how a paragraph is started can strongly impact the sentiment of later internal sentences

II. LITERATURE REVIEW

Linlin You et al [1] The intelligence of Smart Cities (SC) is represented by its ability in collecting, managing, integrating, analyzing and mining multi-source data for valuable insights. In order to harness multi-source data for an informed place design, this paper presents "Public

Sentiments and Activities in Places" multi-source data analysis flow (PSAP) in an Informed Design Platform (IDP). In terms of key contributions, PSAP implements 1) an Interconnected Data Model (IDM) to manage multi- source data independently and integrally, 2) an efficient and effective data mining mechanism based on multi- dimension and multi-measure queries (MMQs), and 3) concurrent data processing cascades with Sentiments in Places Analysis Mechanism (SPAM) and Activities in Places Analysis Mechanism (APAM), to fuse social network data with other data on public sentiment and activity comprehensively. As proved by a holistic evaluation, both SPAM and APAM outperform compared methods. Specifically, SPAM improves its classification accuracy gradually and significantly from 72.37% to about 85% within 9 crowd-calibration cycles, and APAM with an ensemble classifier achieves the highest precision of 92.13%, which is approximately 13% higher than the second best method. Finally, by applying MMQs on "Sentiment & Activity Linked Data", various place design insights of our test bed are mined to improve its livability.

Koyel Chakraborty, , Siddhartha Bhattacharyy and Rajib Bag , et al. [2] In the current era of automation, machines are constantly being channelized to provide accurate interpretationsof what people express on social media.

The techniques of communalizing user data have also been surveyed in this article. Data, in its different forms, have also been analyzed and presented as a part of survey in this article. Other than this, the methods of evaluating sentiments have been studied, categorized, and compared, and the limitations exposedin the hope that this shall provide scope for better research in the future.

Namrata Godbole, Manjunath Srinivasaiah , Steven Skiena , et al. [3] Newspapers and blogs express opinion of news entities (people, places, things) while reporting on recent events. We present a system that assigns scores indicating positive or negative opinion to each distinct entity in the text corpus. Our system consists of a sentiment identification phase, which associates expressed opinions with each relevant entity, and a sentiment aggregation and scoring phase, which scores each entity relative to others in the same class. Finally, we evaluate the significance of our scoring techniques over large corpus of news and blog

Devika M D, Sunitha C, Amal Ganesh,et al. [4] Sentiment analysis (SA) is an intellectual process of extricating user's feelings and emotions. It is one of the pursued field of Natural Language Processing (NLP). The evolution of Internet based applications has steered massive amount of personalized reviews for various related information on the Web. These reviews exist in different forms like social Medias, blogs,Wiki or forum websites. Both travelers and customers find the information in these reviews to be beneficial for their understanding and planning processes. The boom of search engines like Yahoo and Google has flooded users with copious amount of relevant reviews about specific destinations, which is still beyond human comprehension. Sentiment Analysis poses as a powerful tool for users to extract the needful information, as well as to aggregate the collective sentiments of the reviews. Several methods have

come to the limelight in recent years for accomplishing this task. In this paper we compare the various techniques used for Sentiment Analysis by analyzing various methodologies

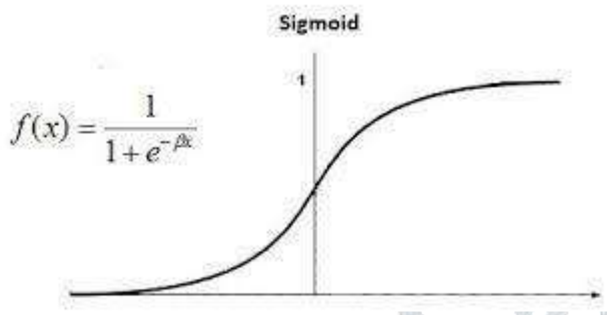
III. METHEDOLOGY

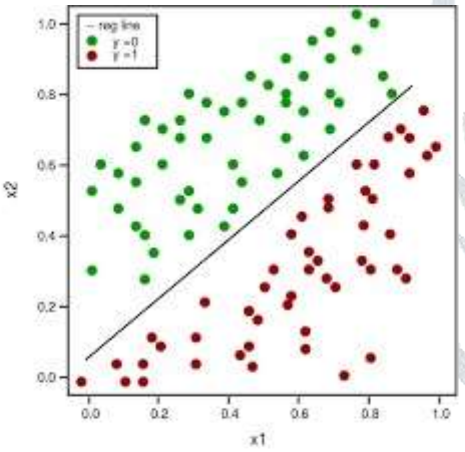
They are divided into six categories which are (SA,ED, SC, FS, TL and BR). The BR category can be classified to lexica, Corpora or dictionaries. The authors categorized the articles that solve the Sentiment classification problem as SC. Other articles that solve a general Sentiment Analysis problem are categorized as SA. The articles that give contribution in the feature selection phase are categorized as FS. Then the authors categorized the articles that represent the SA related fields like Emotion Detection (ED), Building Resource (BR) and Transfer Learning (TL). The fourth column specifies whether the article is domainoriented by means of Yes/No answers (Y or N). Domain-oriented means that domain-specific data are used in the SA process. The fifth column shows the algorithms used, and specifies their categories as shown in Fig. 2. Some articles use different algorithms other than the SC techniques which are presented in Section 4. This applies, for example, to the work presented by Steinberger [43]. In this case, the algorithm name only is written. The sixth column specifies whether the article uses SA techniques for general Analysis of Text

(G) or solves the problem of binary classification (Positive/Negative). The seventh column illustrates the scope of the data used for evaluating the article's algorithms. The data could be reviews, news articles, web pages, micro-blogs and others. The eighth column specifies the benchmark data set or the well-known data source used if available; as some articles do not give that information. This could help the reader if he is interested in a certain scope of data. The last column specifies if any other languages other than English are analyzed in the article. The survey methodology is as follows: brief explanation to the famous FS and SC algorithms representing some related fields to SA are discussed. Then the contribution of these articles to these algorithms is presented illustrating how they

use these algorithms to solve special problems in SA. The main target of this survey is to present a unique categorization for these SA related articles.

Logistic Regression is a supervised learning algorithm that is used when the target variable is categorical. Hypothetical function h(x) of linear regression predicts unbounded values. But in the case of Logistic Regression, where the target variable is categorical we have to strict the range of predicted values. Consider a classification problem, where we need to classify whether an email is a spam or not. So, the hypothetical function of linear regression could not be used here to predict as it predicts unbound values, but we have to predict either 0 or 1.





To do, so we apply the sigmoid activation function on the hypothetical function of linear regression. So the resultant hypothetical function for logistic regression is given below :

h(x) = sigmoid(wx + b)

Here, w is the weight vector.
x is the feature vector.
b is the bias.

Sigmoid (z) = 1 / (1 + e(- z))

Mathematical Intuition:

The cost function of linear regression (or mean square error) can't be used in logistic regression because it is a non-convex function of weights. Optimizing algorithms like i.e gradient descent only converge convex function into a global minimum.
So, the simplified cost function we use :

J = - ylog(h(x)) - (1 - y)log(1 - h(x))
here, y is the real target value
h(x) = sigmoid(wx + b)
For y = 0,
J = - log(1 - h(x))
and y = 1,
J = - log(h(x))

This cost function is because when we train, we need to maximize the probability by minimizing the loss function

Gradient Descent Calculation:

```
repeat until convergence {
      tmpi = wi - alpha * dwi
      wi = tmpi
}
```

where alpha is the learning rate.
The chain rule is used to calculate the gradients like i.e dw.

Chain rule for dw

here, a = sigmoid(z) and z = wx + b.

IV. RESULTS & DISCUSSION

We used the twitter dataset publicly made available by Stanford university. Analyses was done on this labeled datasets using various feature extraction technique. We used the framework where the preprocessor is applied to the raw sentences which make it more appropriate to understand.(means sentences is cut into words , remove the stopwords) Further, the different machine learning techniques trains the dataset with feature vectors and then the semantic analysis offers a large set of synonyms and similarity which provides the polarity of the content.

Dataset Description :

Train Data	45000
Negative	23514
Positive	21486

Test Data	44832
Negative	22606
Positive	22841

Fig. 3. Project windows and real time results

V. CONCLUSION

Nowadays, sentiment analysis or opinion mining is a important topic in machine learning. We are still far to detect the sentiments of s corpus of texts very accurately because of the complexity in the English language and even more if we consider other languages such as Chinese. In this project we tried to show the basic way of classifying comments into positive or negative category using Logistic Regression . We could further improve our classifier by trying to extract more features from the comments, trying different kinds of features, tuning the parameters of the Logistic Regression classifier, or trying another classifier all together.

We presented a new approach towards reviewing of online learning contents.

We used a logistic regression algorithm which is machine learning algorithm and improve accuracy.

Complex sarcastic comments can be analyzed.

Rating of blog is generated based on which user can decide to wheather to read a blog or not.

VI. REFERENCES

1. "Koyel Chakraborty , Siddhartha Bhattacharyy and Rajib Bag A Survey of Sentiment Analysis from Social Media Data IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS"

2. "Harnessing Multi-source Data about Public Sentiments and Activities for Informed Design Linlin You, Member, IEEE, Bige Tunc¸er, Member, IEEE, and Hexu Xing IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING "

3. Namrata Godbole,Manjunath Srinivasaiah,Steven Skiena :"Large-Scale Sentiment Analysis for News and Blogs

4. Devika,MD,Sunitha,Amal Ganesh :" Sentiment Analysis :A Comparative Study on different approaches."

5. M. Hoffman, D. Steinley, K. M. Gates, M. J. Prinstein, and M. J. Brusco, "Detecting clusters/communities in social networks," Multivariate Behav. Res., vol. 53, no. 1, pp. 57–73, 2018, doi: 10.1080/ 00273171.2017.1391682

6. J. Leskovec, "Social media analytics: Tracking, modeling and predict- ing the flow of information through networks," in *Proc. 20th Int. Conf. Companion World Wide Web*, Mar. 2011, pp. 277–278.

7. F. Neri, C. Aliprandi, F. Capeci, M. Cuadros, and T. By, "Sentiment analysis on social media," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2012, pp. 919–926.

8. C. Bron and J. Kerbosch, "Algorithm 457: Finding all cliques of an undirected graph," *Commun. ACM*, vol. 16, no. 9, pp. 575–577, 1973.

9. R. L. Breiger, S. A. Boorman, and P. Arabie, "An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling," *J. Math. Psychol.*, vol. 12, no. 3, pp. 328–383, 1975.

10. S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, nos. 1–7, pp. 107–117, 1998.

11. J. M. Kleinberg, "Authoritative sources in a hyperlinked environment,"

J. ACM, vol. 46, no. 5, pp. 604–632, 1999.

12. C. Canali, M. Colajanni, and R. Lancellotti, "Data acquisition in social networks: Issues and proposals," in *Proc. Int. Workshop Services Open Sources (SOS)*, Jun. 2011, pp. 1–12.

13. B. Wellman, "The development of social network analysis: A study in the sociology of science," *Contemp. Sociol.*, vol. 37, no. 3, p. 221, 2008.

14. D. Jensen and J. Neville, "Data mining in social networks," in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers* (Computer Science Department Faculty Publication Series). Amherst, MA, USA: Univ. of Massachusetts, 2003, pp. 287–302.

15. C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom, "Expertise identification using email communications," in *Proc. 12th Int. Conf. Inf. Knowl. Manage.*, Nov. 2003, pp. 528–531.

16. J. Neville and D. Jensen, "Collective classification with relational dependency networks," in *Proc. Workshop Multi-Relational Data Min- ing (MRDM)*, 2003, p. 77.