# The Comparison Between Lexicon-Based and Machine Learning In Sentimental Analysis

Misha Jain[1], Dr. B. K. Verma[2]
*Department of Computer Science, CEC Landran, Mohali*

***Abstract -*** For the fast-growing environment, there is a need to analyze sentiments with accurate results and limited time consumption. Hadoop architecture uses sentimental data for processing. As it needs a large amount of data for analysis so it requires techniques that can give a fast and accurate result. At present, word sentimental orientation identification researchers mainly fall into two categories: Machine learning and semantic comprehension, machine learning seems to work in specific field words, but cannot handle general-field words effectively and semantic comprehension also cannot get ideal scores at precision and recall, therefore, we put forward a fusion of trans-dative learning and semantic comprehension for determining words' sentimental orientation. The sentimental analysis uses tree data structure. This data approach uses systematic way for traversal. This data, with studies of the bloom filter, can grow at a high rate efficiently with the use of the suitable hash function. Sentimental analysis needs a large data for complete analysis, so the Hadoop framework can provide the platform to store large dataset and process it.

***Keywords -*** Data collection, Sentiment detection, Presentation of the output, Supervised Machine learning based techniques, Lexicon Based, Hybrid Techniques

## I. INTRODUCTION

Sentimental analysis has emerged as a state of art domain with significant contributions from both industry and research community. High diversity of data resources and textual disorder are the primary reasons behind this idea. Automatic opinion recovery and summarization tasks are target conceptual areas that require an extensive study of sentiments, sentiments, and opinions expressed in textual form over the network [1]. Explosive enhancement of social media content, e-business, rating systems, online forums, and businesses are add-ons for the vast data resource. Hence, subjecting related sentences with meaningful opinions, reading and summarizing them into a usable form need an interface of automated opinion discovery and summarization tools. Sentiments are usually identified as either positive and negative opinions or emotions. Sentimental analysis often comprises of terms such as opinion mining, appraisal extraction from structured and unstructured documents. [2] Also, it is relates to text mining, computational linguistics and natural language processing in technical aspects Sentimental Analysis can also be used for identification of sarcastic tweets, determining polarity through it and predicting the results of some political results up to an efficient level. [3] Automatic means of sentimental analysis leads to the concept of polarity, being marked by the words according to their semantic orientation. It is usually termed as prior polarity and contextual polarity where a certain word's instance can exhibit different polarity.

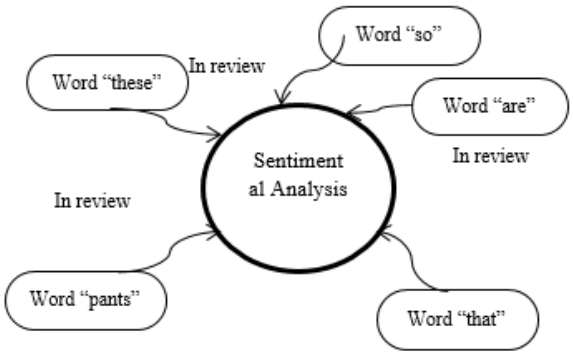


Fig.1: Sentimental analysis

The sentiment initiates within comments, feedback or critiques provide useful indicators for many dissimilar purposes. These opinions can be categorized either into two categories: positive and negative; or into a n-point gauge, e.g., very decent, good, acceptable, bad, very bad. It's also some advantages like the ability to adapt and create trained models for specific purposes and contexts and wider term coverage. It's also Lexicon/learning symbiosis, the detection and dimension of sentiment at the impression level and the lesser sensitivity to changes in the topic domain. Some drawback of sentimental is low applicability to new data because it is necessary the availability of labeled data that could be costly or even prohibitive. A finite digit of words in the dictionaries and the meeting of a fixed sentiment orientation score two words and Noisy reviews [9].

## II. RELATED WORK

Tri Doan et.al (2016) [4]present a variant of online random forests to perform sentiment analysis on customers' reviews. Our model is able to achieve accuracy similar to offline methods and comparable to other online models.Oscar Araque et.al(2016)[5] discribe a hybrid model consisting of a word embeddings model used in conjunction with semantic similarity measures in order to develop an aspect classifier

module. Second, we extend the context detection algorithm by Mukherjee et al. to improve its performance . Lukasz Culer et.al (2016) [6] present adapt a few sentiment analysis methods to obtain sentiment value for social network statements in the Polish language. Developed methods represent different approaches. Starting with PMI-IR, expansion of dictionary through conjunctive connections method, to determining bigram sentiment by analyzing its neighborhood Soujanya Poria (2016)[7] present a novel method to extract features from visual and textual modalities using deep convolutional neural networks. By feeding such features to a multiple kernel learning classifier, we significantly outperform the state of the art of multimodal emotion recognition and sentiment analysis on different datasets.

## III. SENTIMENTAL ANALYSIS PROCESS

A graphical description of the processes is involves in sentiment analysis is detailed in Figure 2 below.

### *A. Data collection*

Sentiment analysis takes advantage of the vast user-generated content over the internet. The data source opinions to queries of user discussions on public forums like blogs, discussion boards, and product review boards as well as on private logs by social network sites like Twitter and Facebook. Very often, the data log is bulky, disorganize sized, and disintegrated on multiple portals. Opinions and feelings are expressed in different ways, including a number of details given, type of vocabulary used, the context of writing, slangs and lingua variations are just a few examples [8].

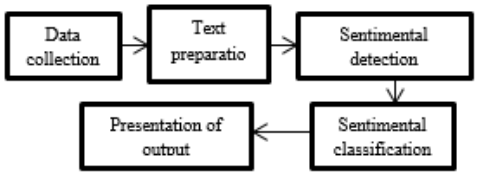


Fig.2: Sentiment analysis process

### *B. Text preparation*

Text preparation involves cleaning the extracted data before the analysis is executed. Usually, text preparation involves identifying and eliminating contextual content from the text dataset, and any information that can reveal the identities of reviewers including reviewer name, reviewer location, review date. In the calculation, any other content that is not deemed relevant to the area of study is also removed from the written data set such as includes stop words or words that are not relevant to the course of analysis.

### *C. Sentiment detection*

The third step is sentiment detection. Sentiment detection requires appraising and extracting reviews and opinions from the textual dataset by the use of computational tasks. Each sentence is examined for subjectivity. Only sentences with subjective expressions are reserved in the dataset. Sentences that convey facts and objective communication are discarded from further seven analyses. Sentiment detection is done at different levels either single term, phrases, complete sentences or complete document with normally used techniques.

### *D. Sentiment classification*

The fourth stage is polarity classification which classifies every subjective sentence in the text dataset into classification groups. Mainly these groups are represented on two extreme points on a continuum (positive, negative; good, bad; like-dislike). However, classification can also involve many points similar to the star ratings used by hotels, restaurants, and retailers.

### *E. Presentation of output*

The general purpose of the analysis is to convert unstructured fragmented text into meaningful info. Once the analysis is completed, a number of conventional options are used to display the result of text analysis. Chief amongst them is the use of graphical displays such as pie charts, bar charts, and line graphs. The split is segmented on color, frequencies, percentages and size. The format of the presentation depends on the research interest.

## IV. SENTIMENT ANALYSIS TECHNIQUES

Sentiment Analysis can be performed in three ways:-

a) Sentiment Analysis based on Supervised Machine learning method
b) Sentiment Analysis by using Lexicon based Technique and
c) Sentiment Analysis by combining the above two approaches.

### *A. Supervised Machine learning based Techniques*

In Supervised Machine learning methods, two types of data sets are required: training data set and test data set. An automatic classifier learns the classification truth of the document from the training set and the accuracy in classification can be evaluated using the test set.

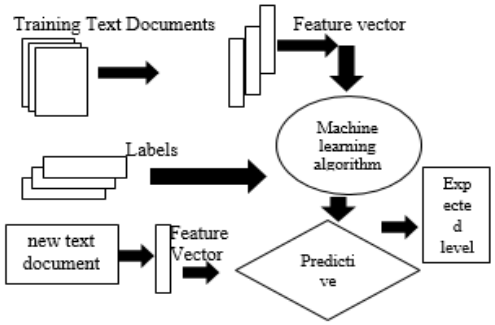


Fig.3: Supervised learning summary

*B. Lexicon Based Technique*

Lexicon Based Technique is an Unsupervised Learning approach since it does no need prior training data sets. It is a semantic orientation approach to belief, mining in which sentiment polarity of benefit present in the given document is determined by relating these features with semantic lexicons. Semantic lexicon comprises lists of the word whose sentimentality orientation is determined already. It classifies the file by aggregating the sentiment alignment of every opinion words present in the document, documents with more positive word lexicons is categorized as positive document and the documents with more negative word lexicons are classified as a negative document.

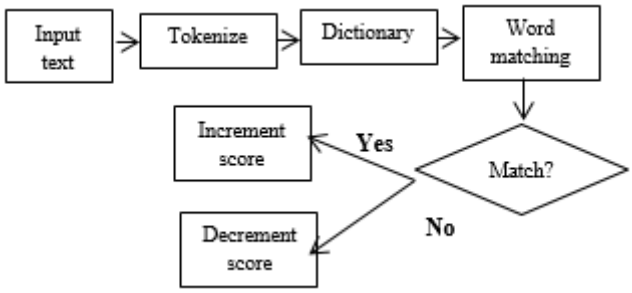


Fig. 4: Lexicon Technique

*C. Hybrid Techniques*

In Hybrid Techniques both mixtures of machine learning and lexicon base approaches ate castoff. Researchers have proved that this mixture gives an improved performance of classification. Minas et al.[10] proposed an idea level sentiment analysis system, called pSenti, which is developed by combining lexicon based and learning-based approaches. The main benefit of their hybrid approach using a lexicon/learning symbiosis is to find the best of together worlds- stability as well as readability from a carefully planned lexicon, and the high accurateness from a great supervised learning algorithm.

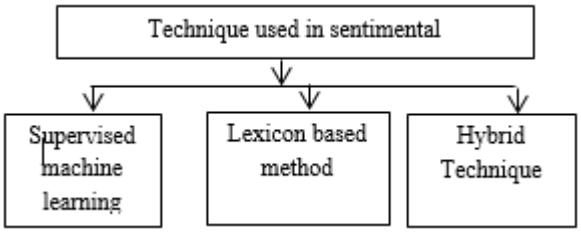


Fig.5: lexicon Technique

Table 1. Comparison between Supervised and Unsupervised techniques

| Supervised | Unsupervised |
|---|---|
| 1. One or more layers of hidden neurons that are not part of the input or output layers of the network that enable the network to learn and solve any complex problems<br><br>2. The nonlinearity reflected in the neuronal activity is differentiable.<br><br>3. The interconnection model of the network exhibits a high degree of connectivity | 1. It transforms an incoming signal pattern of arbitrary dimension into one or 2 dimensional map and perform this transformation adaptively<br><br>2. The network represents a feedforward structure with a single computational layer consisting of neurons arranged in rows and columns.<br><br>3. At each stage of representation, each input signal is kept in its proper context and, neurons dealing with closely related pieces of information are close together and they communicate through synaptic connections. |

Table 2 Techniques Used in Sentimental Analysis

| Technique Name | Merits | Demerits |
|---|---|---|
| Naïve Bayes | Probability,Easy to Implement | Less Accuracy |
| Support Vector Machine | Risk minimize | Speed and Time |
| K-NN | Sufficiently large set | Lazy Learner |
| ANN | Follows the theoretical aspects, improve accuracy | hard to debug |
| Lexicon Based Approach | Evaluates sement polarity for review | cannot be divided into different units in real life |
| Rule-Based Approach | No training phase is required | fails to ascertain the polarity of the text |

Table 3: Comparative Study of Sentiment Analysis

| Approach | Dataset | Approach |
|---|---|---|
| Supervised [11] | Movie Review | NB,SVM,ME |
| Supervised[12] | Movie Review | SVM |
| Unsupervised[13] | Movie review bank and automobile | PMI |
| Unsupervised[14] | Moview Review | Lexicon |
| Hybrid[15] | Twitter Tweets | ML and Lexicon |
| Hybrid[16] | Multi-domain | ML and Lexicon |

## V. CONCLUSION

In this paper, it discussed about full sentimental process then we discuss about its techniques use in sentimental process mainly three types of techniques are used in this paper:

1) Sentiment Analysis based on Supervised Machine learning technique,
2) Sentiment Analysis by using Lexicon based Technique and
3) Sentiment Analysis By combining the above two approaches .We discuss the full working process of sentimental analysis with all parts and working each of them.

A lot has been researched in this field, but still, there are many issues as sentiment analysis processes text based unstructured data. A Dictionary based approach takes less processing time than supervised learning approach, but accuracy is not up to the mark. Supervised learning approach provides better accuracy. From this survey, it can be concluded that supervised techniques provide better accuracy compared to dictionary based approach.

## VI. REFERENCES


[1]. Bing Liu, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies 5.1, pp. 1-167, 2012.

[2]. Nasukawa, Tetsuya, and Jeonghee Yi, "Sentiment analysis: Capturing favorability using natural language processing," Proceedings of the 2nd international conference on Knowledge capture, ACM, 2003.

[3]. Tayal, Devendra Kr, et al., "Polarity detection of sarcastic political tweets," International Conference on Computing for SustainableGlobal Development (INDIACom), IEEE, 2014.

[4]. Doan, Tri, and Jugal Kalita. "Sentiment Analysis of Restaurant Reviews on Yelp with Incremental Learning."

[5]. Araque, Oscar, Ganggao Zhu, Manuel Garcıa-Amado, and Carlos A. Iglesias. "Mining the Opinionated Web: Classification and Detection of Aspect Contexts for Aspect Based Sentiment Analysis."

[6]. Culer, Łukasz, and Olgierd Unold. "Sentiment Analysis of Social Networks Statements for the Polish Language." In Network Intelligence Conference (ENIC), 2016 Third European, pp. 134-139. IEEE, 2016.

[7]. Poria, Soujanya, Iti Chaturvedi, Erik Cambria, and Amir Hussain. "Convolutional MKL based multimodal emotion recognition and sentiment analysis." ICDM, Barcelona (2016).

[8]. Rambocas, Meena, and João Gama. Marketing research: The role of sentiment analysis. No. 489. Universidade do Porto, Faculdade de Economia do Porto, 2013.

[9]. Alessia, D., et al. "Approaches, Tools and Applications for Sentiment Analysis Implementation. " International Journal of Computer Applications125.3 (2015).

[10]. Mudinas, Andrius, Dell Zhang, and Mark Levene. "Combining lexicon and learning based approaches for concept-level sentiment analysis." In Proceedings of the first international workshop on issues of sentiment discovery and opinion mining, p. 5. ACM, 2012.

[11]. B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135.

[12]. A. Abbasi, H. Chen and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," In ACM Transactions on Information Systems, vol. 26 Issue 3, pp. 1-34, 2008.

[13]. P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", Proceedings of the Association for Computational Linguistics (ACL), 2002, pp. 417–424.

[14]. A. Harb, M. Planti, G. Dray, M. Roche, Fran, O. Trousset and P. Poncelet, "Web opinion mining: how to extract opinions from blogs?", presented at the Proceedings of the 5th international conference on Soft computing as trans-disciplinary science and technology, Cergy-Pontoise, France, 2008.

[15]. L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", Technical report, HP Laboratories, 2011.

[16]. Ji Fang and Bi Chen, "Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classification", In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), pages 94–100, 2011.