Sentiment analysis on micro-blog data using machine learning techniques - A Review

M S Kalaivani¹, S Jayalakshmi²
Tagore College of Arts and Science, Chennai,600049,TamilNadu,India
VELS University,Chennai, 600117, Tamilnadu,India
Email: kalaims2007@gmail.com,jai.scs@velsuniv.ac.in

Abstract :The user generated content in social networks, websites, ratings and reviews about product, people, places have become very beneficial for various fields to improve its growth. Sentiment analysis, which studies user generated content and find the correct sentiment polarity. Data collection, feature extraction, sentiment classification and decision making are the steps involved in this process. Machine Learning and Deep learning Methods are used to predict the sentiments. Large Input dataset from websites or microblog is taken as input and processed. There are some challenges in identifying accurate decisions. This Review Paper focus on recent studies about learning models such as Support vector Machines ,Maximum Entropy , deep learning models and various methods which are used in solving sentiment classification, product review analysis, Text Processing and visual analysis etc.,

Keywords : sentiment analysis, Machine Learning ,Deep Learning models, Support vector Machine, Convolutional Neural networks.

1. Introduction
Nowadays, enormous amount of information is presented in webpages, databases, and blogs as online document. The unstructured form of data from the internet required to be changed well-structured data, to analyze further. The Analysis process used in many fields such as Consumer information, websites, and Social Media [1]. The goal of sentiment analysis is study user feedback and predict the sentiments. Usually, It defines opinion polarity, i.e., if somebody has positive, negative, or neutral opinion towards something. For example, (i) if a person plan for a trip to a particular place, he simply refer the online feedback of previous visitors for making the conclusion.
Supervised learning
Machine learning
UnSupervised learning
Reinforcement learning
Supervised learning
Hierarchical C-Means K-means, etc
Classifier
Deep learning
Statistical
Ensemble Methods
Structural
Regression:
- Linear
- Logistic
- Multiple
Naive Bayesian networks
Bayesian networks
Rule-Based: Decision trees
Distance Based SVM KNN
Neural Networks
CNN
RNN
DBN, DNN

Figure 1. Process of sentiment Analysis

(ii) If a person want to purchase a product, he reads online ratings and comments first, and take a decision to buy the product or not. The need of sentiment analysis is raised because requirement of analyzing and formatting the unstructured data in social media or websites.[2]. The idea of sentiment analysis is, without human intervention it determines the way of user reviews [3].
In social media, Twitter has made a big revolution, which is connecting millions of people worldwide through micro blogging service. The registered user of twitter can post short messages about anything, these short messages are called Tweets, which can have character-set length upto 280. In the recent years researchers applied SA techniques with twitter comments and the results helpful in decision making. Machine learning techniques are applied by extracting the sentences and features. The polarity score has been assigned to the lexicons and sum of the polarity score is calculated from the overall values. [4]

1.1. Sentiment analysis with product review

The human’s opinions, sentiments towards a product are being analyzed by the machine and the results encourage the purchasing decision of the customers. Product review analysis is an entirely recognized field, where the customers can easily decide their product needs. Initially the product reviews will be taken as input text, and preprocessing is done. In this phase the input file is normalized, which ready for training process. This file is mapped with the saved features for training.

After mapping the data will be fed into the different models and the performance of the models are analyzed as results. Several Machine Learning and are proposed in the past. Deep learning models, a division of Machine learning well suited for sentiment analysis with large dataset. The following is the basic procedure in sentiment analysis for product review.

Procedure for SA process

Step 1: Input text paragraph from dataset
Step 2: Preprocessing the text. It includes removing stop words, removing numbers, expanding acronyms, replacing negation and removing URLs.
Step 3: Feature Extraction phase: Extracting words corresponds to adjective, adverb and verb.
Step 4: Selection of essential features for training the input data. Mapping the input data with selected features.
Step 5: The dataset can be divided 70% training and 30% testing.
Step 6: Training dataset to Machine learning algorithms or train the model with Deep learning techniques
Step 7: Test the model with different values.
Step 8: Result analysis with performance of the model.

2. Literature Review

In this section various studies about sentiment analysis by using Machine Learning and Deep Learning models are explained. Naïve Bayes, Maximum Entropy, Ensemble methods, SVM, Decision trees, Stochastic Gradient descent are some of the familiar machine learning methods used in recent researches.

2.1. Machine Learning Techniques

The Author [5, Monisha Kanakaraj], analyzed Naïves Bayes, Maximum Entropy, Support vector Machines algorithms in his work. Twitter dataset 7086 reviews from election dataset, 25000 reviews from movie dataset were taken as input for sentiment analysis. Randomized Trees classification method produces better outcomes. Author [6, Ana Valdivia], applied sentiment analysis with Trip advisor; it is an American travel website company that provides user opinions about their hotels, restaurant and monuments. The analysis task categorized into three steps, polarity detection, Aspect selection, and classification. The author applied SAMs, SentiStrength, Bing, Syuzhet, and CoreNLP methods to analyze the user opinions. The result shows that Syuzhet detect at best 57.48 and 45.10 percent of negative feedback. In this paper [7, (Geetika Gautam)] Author proposed machine learning techniques for sentence classification and taken product reviews as input dataset from twitter data. The naïve byes method provides better result than maximum entropy. SVM provides improved results in unigram model. The accuracy level is gained in semantic analysis word net to to 89.9% from 88.2%.
In this study [8, Neethu M S] Author, Proposed sentiment analysis on electronic products like mobile, laptops by using machine learning algorithms. They presented a new feature vector for classifying the reviews into positive and negative. User reviews about electronic products taken from twitter as input dataset and performed sentence level sentiment analysis. Naive Bayes, SVM, Maximum Entropy Classifier and Ensemble classifiers are applied and performance analysed. The result shows Naive Bayes has 89.5% accuracy whereas remaining three methods obtained 90% accuracy level. The Author [9, Venkata Sasank Pagolu] predicts the stock market movement of a company by using supervised machine learning algorithms. The classifier Word 2vec representations of tweets trained on Random Forest algorithm. The result shows 70.2% of accuracy with Random forest algorithms and with N-gram classifier it shows the accuracy level up to 70.5%.

In this study [10, Sankar] has proposed sentiment analysis about movie review by taking inputs from thee datasets Polarity, IMDb, and Rotten Tomato and applied the CNN approach in a mobile environment. The result shows that GloVe word embedding performs better. The GloVe 6B and Glove Crawl carry out in the same way. The SSWE word vector not produced expected outcome. In classification part, CNN and trained word vectors works better in mobile applications.

Author [11, Akshi Kumar] specified approach for sentiment analysis on twitter data. User opinions extracted with adjectives, verbs, adverbs from tweets. To get the semantic orientation of adjectives, corpus-based approach was applied. It has been considered a motivational technique for upcoming researches. In this work, Author [12, Akshat Bakliwal] used NLP techniques N-Gram, POS-Tagged features for sentiment analysis. Movie reviews and product reviews are taken as input dataset. Multi-Domain Sentiment Dataset (Version 2.0) has product reviews taken from Amazon.com with many (total 25) categories. They considered 4000 positive and 4000 negative reviews. Movie Review Dataset has 1000 positive and 1000 negative reviews. Machine learning algorithms SVM, MLP and NB applied and results were analysed. In the N-gram technique SVM achieved better results in both datasets 81.15% and 79.4%n POS-tagged approach MLP achieved 81.60% and 79.47% for movie reviews and product reviews.

In this study [13, Chris Jefferson] proposed fuzzy rule based system for sentiment analysis. Fuzzy membership degrees provide more refined outputs. Decision trees and Naivebayes methods compared with fuzzy rule based system. All classifiers achieved good results with accuracy level 0.9. The fuzzy rule based system obtained better performance than other classifier. In this study the author [14, A. A. Aliane)] proposed a genetic algorithm for Arabic sentiment analysis by taking the input dataset from book reviews. They have performed five basic algorithms SVM, Multinomial Naive Bayes (MNB), Naive Bayes (NB), Stochastic Gradient descent, Decision Trees and the outcomes were compared. The research shows that proposed algorithm performs better than previous classifiers. MNB obtained 94% accuracy level with unigrams.

In this study author [15, Zhao Jianqiang1)] evaluated various pre-processing methods on five different input dataset. To identify the sentiment polarity two feature model- Word n-grams features model and Prior polarity score feature model were used. The classifiers SVM, NB, LR, RF, and baseline method (C-Method) were applied. The result shows that, NB and RF classifiers are more efficient than LR and SVM classifiers.
### Table 1: Analysis of Machine Learning Techniques in sentiment analysis.

| Researcher Name and Year | Model Used | Objective | Data Set | Results | Challenges for Future work |
|--------------------------|------------|-----------|----------|---------|-----------------------------|
| Monisha Kanakaraj and Ram Mohana Reddy Guddeti[5] | Ensemble methods and machine learning methods | Sentiment analysis on movie and election reviews | Twitter dataset | Ensemble methods performed 3-5% better than other methods | Dataset used for training and testing can be in different ratio with large subsets, which is randomly selected. |
| Valdivia, M. Victoria Luzón, and Francisco Herrera[6] | SentiStrength, Bing, Syuzhet, and CoreNLP Methods | sentiment analysis with Trip advisor | Dataset taken from American travel website | Syuzhet detects at best 57.48 and 45.10 percent of negative reviews | The values are varying in user ratings and Evaluations of SAM. A new approach is required for efficient classification. |
| Geetika Gautam, Dikvakar Yadav[7] | Machine learning techniques with semantic analysis | Sentiment analysis on Product review | Twitter dataset | The accuracy level is gained in semantic analysis word net to 89.9% from 88.2% | The training data can be enlarged for better results in sentence identification process. |
| Neethu M S, Rajasree R[8] | Naive Bayes, SVM, Maximum Entropy Classifier and Ensemble classifiers | Sentiment analysis on electronic products like mobile, laptops | Twitter dataset | Naive Bayes has 89.5% accuracy whereas remaining three methods obtained 90% accuracy level | The input dataset can be can be increased, All these classifiers have almost similar accuracy for the new feature vector. |
| A. A. Aliane H. Aliane, M. Ziane N. Bensaouli[14] | SVM, Multinomial Naive Bayes (MNB), Naive Bayes (NB), Stochastic Gradient descent Decision Trees | a genetic algorithm proposed for book reviews feature selection in Arabic sentiment analysis | | Proposed algorithm performs better than previous classifiers. MNB obtained 94% accuracy level with unigrams. | Multi domain dataset can be taken as input dataset to apply feature selection based approaches. |
2.2. Deep Learning Techniques

In deep learning, data representation is referred by increasing the abstraction levels [16]. In the previous papers, there are various deep learning Architectures are Discussed. Deep learning is used effectively in both supervised and unsupervised algorithms. Author [17, Islam.J], proposed a CNN framework for predicting the sentiments from social images. This is implemented by using Caffe and Python on a Linux platform. This model can be used in multimedia content to know user feedback, predictive modeling and advertising. A twitter dataset with 1269 images is taken as input and back propagation method is applied. The proposed model achieved better performance than existing systems. AlexNet was used in earlier systems and GoogleNet provides nearly 9% improved performance.

The authors [18, A. Severyn] implemented deep learning methods with twitter data. To initialize the word embedding neural language is used. Embedded words and parameters are applied to initialize the network, and training phase done by using supervised corpus Semeval-2015. Sentence matrix pooling, activations, convolutional layers and softmax are the components used in proposed system. Deep learning technique is implemented in messagelevel, phrase level to predict polarity and better outcomes achieved.

The researcher [19, L. Yanmei] has explained about sentiment analysis with user opinions from a microblog. The user feedback on hot events analyses by using CNN. The corpus which has 1000 micro-blog comment divided into three labels, 274 for neutral, 300 for negative and 426 for positive emotions. The proposed work is compared with earlier models CRF, SVM, other traditional methods and performances analysed. The proposed model achieved the good performance in emotion analysis.

This study presents [20, ZeyuXiong] the CBOW model and CNNs to form a new deep learning model for paragraph vector representation. Stanford Sentiment Treebank dataset is used as input. Amazon Mechanical Turk used generates labels and 10000 paragraphs from Treebank taken as input. The training process needs the 56×50 matrix foreach paragraph. Result shows that CBOW and CNN model achieved better results.

This study [21, Yongping Du] targets at sentiment classification of short messages. The author proposed capsule-based hybrid neural network model to attain the semantic data efficiently. In this model, BGRU (Bidirectional gated recurrent unit) applied to handle interdependent features with long distance. It extracts text to improve quality of the expression. There are different input dataset used in this system. (1) Movie Review Data (2) NLPCC 2014 Data, 10000 product reviews. The performance of the system is compared with attention-based model and CNN (convolutional neural networks) model. The result shows that, movie review data achieved the highest accuracy of 0.8255 than NLPCC2014.

In this study [22, R. Socher], RNTN (Recursive Neural Tensor Network) and Sentiment Treebank model are projected to explain the compositional effects at various sentence levels. In previous models, the meaning of long phrases could not be analyzed well by the tool semantic word spaces. Effective training resources are needed for sentiment detection. The RNTN achieved accuracy level of 80.7% in sentiment prediction. This author [23, T. Chen] developed a sequence model for embedding of reviews of products which has temporal nature. Sentiment analysis is done with machine learning classifiers by taking three datasets as input from Yelp and IMDB. Every review is labeled with respect to rating score. Back-propagation algorithm with Adam stochastic optimization methods are applied in training phase. The result shows that the proposed model performs better in document level sentiment classification.
Table 2: Analysis of Deep Learning Techniques in sentiment analysis.

| Researcher Name | Model Used | Objective | Data Set | Results | Challenges for future work |
|-----------------|------------|-----------|----------|---------|---------------------------|
| Zeyu Xiong[20]  | CBOW model and CNN | paragraph vector representation | 10000 paragraph from Stanford Sentiment Treebank dataset | CBOW and CNN model achieved better results. | Skip-gram and CNNs can be combined to produce paragraph vector. |
| Yongping Du, Xi aoz, heng Zhao, Meng He - Wengyanguo[21] | capsule-based hybrid model | sentiment analysis social of data | Movie Review Data, NLPCC 10000 product reviews | Movie review dataset obtained the highest accuracy of 0.8255. | This model can be extend to aspect-based opinion classification |
| Socher R, Pere- ygin A, and Wu J, [22] | RNTN | Semantic Compositionality | 11,855 sentences from movie review (Pang and Lee 2005) | The RNTN provides 80.7% accuracy with 9.7% over baselines | The training and testing datasets can be increased and various classification methods can be added. |
| Chen T, Xu, R. He, Xia, and Wang , [23] | Recurrent Neural Network(RNN-GRU) | User and Product Distributed Representations | Input dataset from Yelp and IMDB | Proposed model outperformed than other models | sequence learning model, can be explored to bidirectional RNN and LSTM |

3. Conclusion
The need of normalizing the online data increases researches in the field of sentiment analysis. It extracts the unstructured data from social media and examines the user opinions. Machine Learning and Deep Learning Techniques play vital role in pre-processing the texts, classification, training, testing the models and result analysis. The outcomes of the research used to know the user expectations and feedback about a product, place, or any common thing. This analysis provides the way for development in all the fields. In this paper various studies have been discussed about machine Learning and Deep Learning Models which are used in sentiment classification.

4. Analysis with future work
This paper explained about some recent researches in sentiment analysis. There are so many models established by using Machine learning and Deep learning Methods. NaiveBayes, Support Vector Machines. Maximum Entropy, Ensemble methods, Decision trees, Stochastic Gradient descent are most used machine learning algorithms. Convolutional neural networks-CNN, DBN, Recursive neural network - RNN, DNN, RNN are frequently used Deep learning Algorithms. Since Deep Learning models have more hidden layers, performs better than other Techniques. These models provide efficient training in supervised and unsupervised ways and also it performs automatic feature extraction. These models have some challenges, it requires large input dataset for better performance and it is costly
to train. The performance of the algorithm varies depend on the problem and the input dataset. As a future work product review analysis can be done by using Deep learning Methods and the accuracy of the outcomes can be shown better than previous models.

References

[1] Doaa Mohey El-Din Mohamed Hussein, 2016, *Journal of King Saud University - Engineering Sciences* survey on sentiment analysis challenges, Computers and Information Cairo University, Cairo, Egypt.
[2] M. Haenlein and A. M. Kaplan 2010 An empirical analysis of attitudinal and behavioral reactions toward the abandonment of unprofitable customer relationships, *J. Retail.* vol.9, no.4
[3] F. Luo C. Li and Z. Cao, 2016 IEEE 20th Int. Conf. Comput. Support. Coop. Work Des., Affective-feature-based sentiment analysis using SVM classifier, pp. 276281.
[4] Olga Kolchyna, Thairrois T. P. Souza, Philip C. Treleaven and Tomaso18 Sep 2015 Astel12,arXiv:1507.0095v3 [cs.CL] Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination
[5] Monisha Kanakaraj and Ram Mohana Reddy Guddeti 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN) NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers.
[6] Ana Valdivia M. Victoria Luzón, and Francisco Herrera University of Granada, IEEE INTELLIGENT SYSTEMS Sentiment Analysis in TripAdvisor, Published by the IEEE Computer Society.
[7] Geetika Gautam, Divakaryadav, 2012 IJCSI International Journal of Computer Science Issues, Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis, Vol. 9, Issue 4
[8] Neethu M S, Rajasree R, IEEE - 31661, 4th ICCCN 2013 July 4 - 6, 2013, Sentiment Analysis in Twitter using Machine Learning Techniques.
[9] Venkata Sasan, Pagolu, Kamal Nayan Reddy Challap, Ganapatipanda, Babita Majhi, 2016 Intelligent sentiment analysis approach using edge computing-based deep learning technique Sentiment Analysis of Twitter Data for Predicting Stock Market Movements.
[10] H. Sankarl V. Subramaniaswamy1V. Vijayakumar2 Sangiah Arun Kumar3R. Logesh1 A. Umamakeswari1, *Softw: PractExper.* 2019;1–13. wileyonlinelibrary.com, John Wiley & Sons.
[11] Akshi Kumar and Teeja Mary Sebastian, 2012 IJCSI International Journal of Computer Science Issues, Sentiment Analysis on Twitter Vol. 9, Issue 4, No 3
[12] Akshat Bakliwal, Ankit Patil, Priyush Arora, Vasudeva Varma, 2011 Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP Towards Enhanced Opinion Classification using NLP Techniques, pages 101–107.
[13] Chris Jefferson, Han Liu IEEE and Mihaela Cocea 2017 IEEE, Fuzzy Approach for Sentiment Analysis 978-1-5090-6034-4/17/ 2017 IEEE.
[14] Aliane H. Aliane M. Ziane1 N. Bensaou1 2016 IEEE A Genetic Algorithm Feature Selection Based Approach for Arabic Sentiment Classification, 978-1-5090-4320-0/16
[15] Zhao Tianqiang1 Gui Xiaolin 2016 IEEE Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis, 2169-3536
[16] Mohamed-Shamim Kaiser Amir Hussain and Stefano Vassanelli 2018 IEEE Applications of Deep Learning and Reinforcement Learning to Biological Data, 2162-237X
[17] Y. GuoY. Liu A Oerlemans S. Lao S. Wu and M. S. Lew, 2016 Neurocomputing, Deep learning for visual understanding: A review, vol. 187, pp. 2748
[18] J. Islam and Y. Zhang, 2016 IEEE International Conference on Big Data Cloud Comput. (BDCloud), Visual Sentiment Analysis for Social Images Using Transfer Learning Approach, pp. 124130.

[19] A. Severyn A. Moschitti, 2015 Proc. 38th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.-SIGIR15, Twitter Sentiment Analysis with Deep Convolutional Neural Networks, pp. 959962, 2015.

[20] Zeyu Xiong, Qiangqiang Shen, Yijie Wang I and Chenyang Zhu, Paragraph 2018 CMC Vector Representation Based on Word to Vector and CNN Learning, vol. 55, no. 2, pp. 213-227.

[21] Yongping Du, Xiaozheng Zhao, Meng He, Wengyang Guo, 2019 IEEE A Novel Capsule Based Hybrid Neural Network for Sentiment Classification, pp. 2169-3536

[22] R. Socher A. Perelygin, J. Wu, 2013 Recursive deep models for semantic compositionality over a sentiment treebank, Proc., pp. 16311642.

[23] T. Chen, R. Xu, Y. He, Y. Xia, X. Wang, 2016 Learning User and Product Distributed Representations Using a Sequence Model for Sentiment Analysis, no. August, pp. 3444.