iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IX **Month of publication:** September 2020

DOI: https://doi.org/10.22214/ijraset.2020.31647

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IX Sep 2020- Available at www.ijraset.com

Twitter Sentiment analysis using different Algorithms

Shivani Pathak[1], Piyusha Mahajan[2], Ankita Patil[3], Rutuja Patil[4], Mrs. M. M. Swami[5]
[1, 2, 3, 4, 5]*Department of Computer Engineering, All India Shri Shivaji Memorial Society's College of Engineering, Pune*

Abstract: Sentiment analysis is an application of natural language processing. It is also known as emotion extraction or opinion mining. It is a very popular field of research in text mining. The basic idea is to find the polarity of the text and classify it into positive, negative or neutral. Polarity of text is determined from scores identified by VADER. It helps us to understand human decision making or to categorize, analyze and extract opinions from review documents on web sites, blogs, social media, and others in order to understand the consumers. To perform sentiment analysis, there are various algorithms such as SVM, Naïve Bayes and DAN2 which are used to predict the polarities and find their accuracies. There are various tasks like subjectivity detection, sentiment classification, aspect term extraction, feature extraction, keyword selection and keyword analysis etc. that are needed to determine the polarity.

Keywords: Natural language Processing, Sentiment analysis, Machine learning, Feature engineering, DAN2 classifier, SVM, Naïve Bayes

I. INTRODUCTION

Social media is in trend for every generation such as Facebook, Twitter with which anyone can express their opinion regarding any topic. These comments can be analyzed to find the hidden motive or real meaning behind them. This can be achieved by Sentiment Analysis. Sentiment analysis is an application of NLP which can help us in opinion mining or emotion extraction by analyzing polarity of sentences. The subjective expression categorized as positive, negative, neutral is known as polarity of sentences.

Sentiment analysis is proven to be very useful in many streams like application in review system, survey response system, marketing, for analysis of product recommendation, etc. These systems can categorize sentences into polarity, can identify emotion or sentiment (happy, sad) or can mark product (interested or not interested) [11]. The data of various companies related to their services or product is stored in unstructured format, so to analyze such data manually is time-consuming and exhaustive. To save time in such cases, Sentiment Analysis is used to scale the data efficiently which is also cost-effective. It also helps in analyzing situations by identifying critical information from the reviews and then performing action and spreading awareness of the situations in Real-time [11]. Sentiment Analysis can be done in different types in which it can perform analysis of text, sentences or voice.

There are various algorithms used for sentiment analysis such as SVM (Support Vector Machine), Naive Bayes classifier, neural network classifier, etc. SVM and Naive Bayes algorithms can be efficient for small dataset but as the data increases the efficiency of these algorithms may decrease [11]. Hence to overcome these problem neural network classifiers such as DAN2 (Dynamic architectural artificial neural networks) algorithm can be used [10]. It can help in the analysis of huge amounts of data and it also provides scalability. Some classifiers can identify only strong opinions but in my case mild opinions also play vital roles like in brand management or for improvising the reputation of a company. Feature engineering used in sentiment analysis with which mild opinion can be identified. DAN2 classifiers with Feature engineering can be used for training and testing of data by which polarity of both strong and mild opinions and emotions can be identified [11].

II. LITERATURE SURVEY

Mainly the nature of the sentence is divided into two categories: subjective and objective. Sentiment analysis depends on the nature of the sentence, which can be either subjective or objective. The objective statements are based on facts and cannot be used for extracting any emotions or sentiment. For example, a sentence like "McDonalds has many outlets in India", states fact and does not contain any emotions.

The subjective statements can be used for identifying the emotions from documents, sentences or phrases. While understanding the sentiments in a sentence (subjective), first the sentence is pre-processed using different techniques and tagged into different POS (Parts of Speech). Sentiments are classified using two main approaches that are Subjective lexicons and Machine Learning [1].

Lexicon-based approaches extract opinions of sentences on the basis of dictionaries. Hassan Saif, Yulan He, Miriam Fernandez, Harith Alani suggested a lexicon-based approach- SentiCircle, which combines context of words used in the text for analysing the sentiments. The outcomes of SentiCircle are better than the other commonly used lexicon-based approaches. However, sentiment lexicons are time consuming and costly [2].

Supervised machine learning techniques outperforms subjective lexicons technique. Some of the most common classifiers used are SVM (Support Vector Machine) and Naïve Bayes [11]. Naive Bayes classifier is a probabilistic approach which is derived from Bayes theorem. The implementation of Naïve Bayes is very easy but the assumption of strong independence between features is not accurate [3] [11].

The Dual prediction technique uses SVM classifier [4] where the emotions are extracted by finding the average of evaluation of text in both reverse and forward directions.

Another approach uses optimized SVM by adding Radial Basis Kernel function (RBF) to traditional SVM [5].The optimized SVM outperforms traditional SVM and Naïve Bayes. First features of data are extracted efficiently then classifier models are applied. Minimum Redundancy and maximum Relevance is one of the approaches which identifies relationships first and then constructs concepts based on them using ConceptNet source [6].

The automatic keyword selection (AKS) is a feature selection technique which outperforms the mRMR technique. The AKS technique reduces the training time required for classifiers like RBF, MLP (Multilayer Perceptron), Naïve Bayes and Decision Tree. This technique is very efficient for huge training datasets [7] unlike Naive Bayesian technique which works efficiently for smaller datasets.

The most common supervised learning classifiers used for sentiment analysis like SVM and Naïve Bayes are old, slow techniques and require excessive computations apart from classification like selection of a kernel function for SVM. A more recent classifier used for sentiment analysis is the Dynamic Architecture for Artificial Neural Networks (DAN2). Ghiassi, M. and Saidane.H developed the DAN2 model which is a modification of the Artificial neural network. It is a feed-forward technique. DAN2 provides better scalability as compared to previous classifiers. It uses all of the samples for training which reduces the training SSE (Sum of squared errors) or MSE (mean squared error). DAN2 compile results at every stage and the user doesn't have to decide the number of hidden nodes for every hidden layer [8].

The DAN2 classifier can also be used effectively for automated text classification and outperforms SVM and kNN (k-Nearest Neighbour). This is especially because of the dimensionality reduction property of DAN2 which is important due to the complexity of the texts used for classification [9].

Most of the approaches used for sentiment analysis usually use strong sentiments in text. The users having mild sentiments in text can be used for change in opinion. These mild sentiments can be identified by using feature engineering with DAN2. David Zimbra, M. Ghiassi, and Sean Lee [10] presented an approach which combines feature engineering and DAN2 classifier for sentiment analysis [10].

III.PROPOSED SYSTEM

The proposed system will take input from twitter dataset as excel/csv sheet. The data is then pre-processed using NLP (Natural Language Processing) techniques like tokenization, stemming, removal of stop word, etc. This pre-processed data is passed on for its transformation to be used in the classifiers. The polarity of the tweets is identified and used to identify the sentiment of the tweets. For the purpose of sentiment analysis three major algorithms, namely Support Vector Machine (SVM), Naïve Bayes and Dynamic Architecture for Artificial Neural Network (DAN2) are used. The results of these algorithms are compared to find the best suited algorithm.

The diagrammatic representation of our system is given as below in Fig. no. 1.

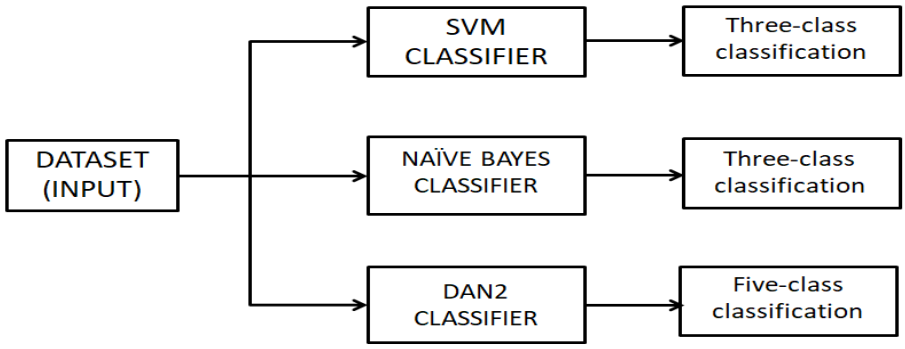


Fig no. 1: The Proposed System

A. Support Vector Machines (SVM)

Support Vector Machines are a popular choice for sentiment analysis. A Support Vector Machine (SVM) is a supervised learning algorithm. It is a discriminative classifier which outputs an optimal hyperplane which categorizes new examples and predicts their classes. In the proposed system, the polarity of the tweets is identified first. Based on this polarity (training data), the SVM classifier is used to predict new samples.

B. Naïve Bayes

Naive Bayes is a family of probabilistic algorithms that take advantage of probability theory and Bayes' Theorem to predict the class of a text. Similar to the approach for SVM, the polarities are identified and the classifier is used to predict new tweets.

C. Dynamic Architecture for Artificial Neural Networks (DAN2)

The original twitter dataset containing airline sentiment tweets is classified into three classes as positive, negative and neutral. But for the objective of fine-grained sentiment classification, five-class classification is required. The **V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner (VADER) is a rule-based library used for sentiment analysis. The VADER library specifically focuses on social media texts. This library can be used to find the sentiment score of the tweets.

Based on the sentiment score, the tweets can be classified into five classes namely strongly negative, weakly negative, neutral, weakly positive and strongly positive. The dataset is then classified using the DAN2 model.

The DAN2 algorithm is a modification of the traditional ANN. It is a purely feed-forward model [11] [10]. DAN2 categories the sentence into 5 types of polarity as mildly positive, strongly positive, neutral, mildly negative, strongly negative. Its architecture is similar to traditional ANN but the number of hidden nodes in hidden layers is fixed in DAN2.

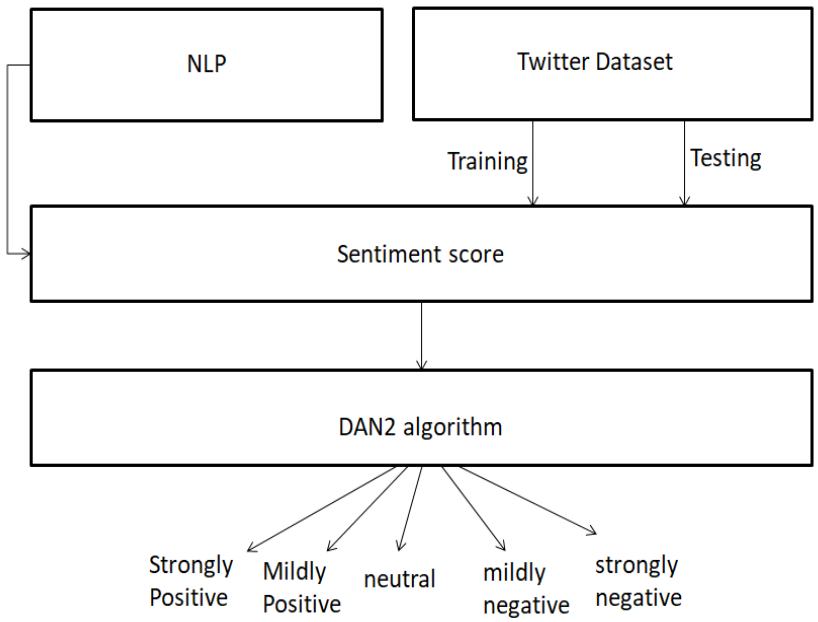


Fig no 2: Implementation of DAN2 for fine-grained sentiment analysis

IV. RESULTS

The Support vector machine classifier model and Naive Bayes classifier model was used to classify the Sentiment of sentences in a dataset depending on the polarity of each sentence.

Initially the data cleaning and normalization was done by using NLTK python library in which removal of stopwords, removal of extra space, removal of extra character like @, #, etc.

After the normalization process, n-grams of sentences were extracted for further process. Sentiment score of every normalized_ tweet is calculated using 'sentiment_score', a textblob library function. So depending on the sentiment_score of every normalized _tweet the polarity was found in 3 categories i.e. positive, neutral and negative. These models were trained and tested. Both models have accuracy rates of 48.63 and 48.64 respectively.

In the Dynamic Architecture of Artificial Neural Network (DAN2) classifier model, data is cleaned using functions like removal of special characters, removal of stop words, removal of extra spaces and replacing short forms of words with real words. This way normalized tweets, tokenization and n-grams of tweets are formed for further evaluation. Python package VADER (Valence Aware Dictionary and Sentiment Reasoner) lexicons based method used for sentiment scoring n-grams of normalized_tweets. DAN2 is a fine-grained Sentiment which divides the polarity into 5 categories strongly positive, strongly negative, weakly positive, weakly negative and neutral based on the sentiment score of VADER. Unlike Naive Bayes and SVM models it decides the polarity in 5 types for mild sentiment analysis with 26.64 percent accuracy.

V. CONCLUSIONS

In this research, we have performed sentiment analysis on SVM, Naïve Bayes and DAN2 algorithms. Three-class classification is performed on the dataset using SVM and Naïve Bayes. For the identification of mild sentiments, five-class classification is performed on DAN2.

As mentioned in section [IV], the accuracy of the DAN2 model is lower than that of SVM and Naïve Bayes. For the DAN2 model, we conducted two experiments for three-class and five-class classification. The accuracy of the DAN2 model is more than the SVM and Naïve Bayes models for three-class classification and lesser for five-class classification. This suggests that the DAN2 model is not well-suited for fine-grained sentiment analysis on the given dataset.

VI. ACKNOWLEDGMENT

We would like to extend our sincere gratitude to our guide Mrs. M.M.Swami for her invaluable guidance and for giving us useful inputs and encouragement time and again, which inspired us to work harder.

We are extremely grateful to Mr. D.P.Gaikwad, Head of Department of Computer Engineering, All India Shri Shivaji Memorial Society's College of Engineering, Pune for his encouragement during the course of the project work.

REFERENCES

[1] Harpreet Kaur, Veenu Mangat, Nidhi, "A Survey of Sentiment Analysis techniques", International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2017).

[2] Hassan Saif, Yulan He, Miriam Fernandez, Harith Alani, "Contextual semantics for semantic analysis of Twitter", Information Processing Management, Volume 52, Issue 1, Jan 2016, ScienceDirect.

[3] Mykhailo Granik and Volodymyr Mesyura, "Fake news detection using naïve Bayes classifier", First Ukraine Conference on Electrical and Computer Engineering(UKRCON), IEEE, 2017.

[4] Minara P Anto, Nivya Johny, Mejo Antony, Vinay James, Muhsina K M, Aswathy Wilson, "Product Rating Using Sentiment Analysis" ,International Conference on Electrical, Electronics and Optimization Techniques(ICEEOT) – 2016.

[5] Bhumika M. Jadav, Vimalkumar B. Vaghela, "Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis", International Journal of Computer Applications, Volume 146 - No. 13,2016.

[6] Basant Agarwal, Soujanya Poria, Namita Mittal, Alexander Gelbukh, Amir Hussain, "Concept-Level Sentiment Analysis with Dependency-Based Semantic Parsing: A Novel Approach, Volume 7, Issue 4, August 2015, Springer.

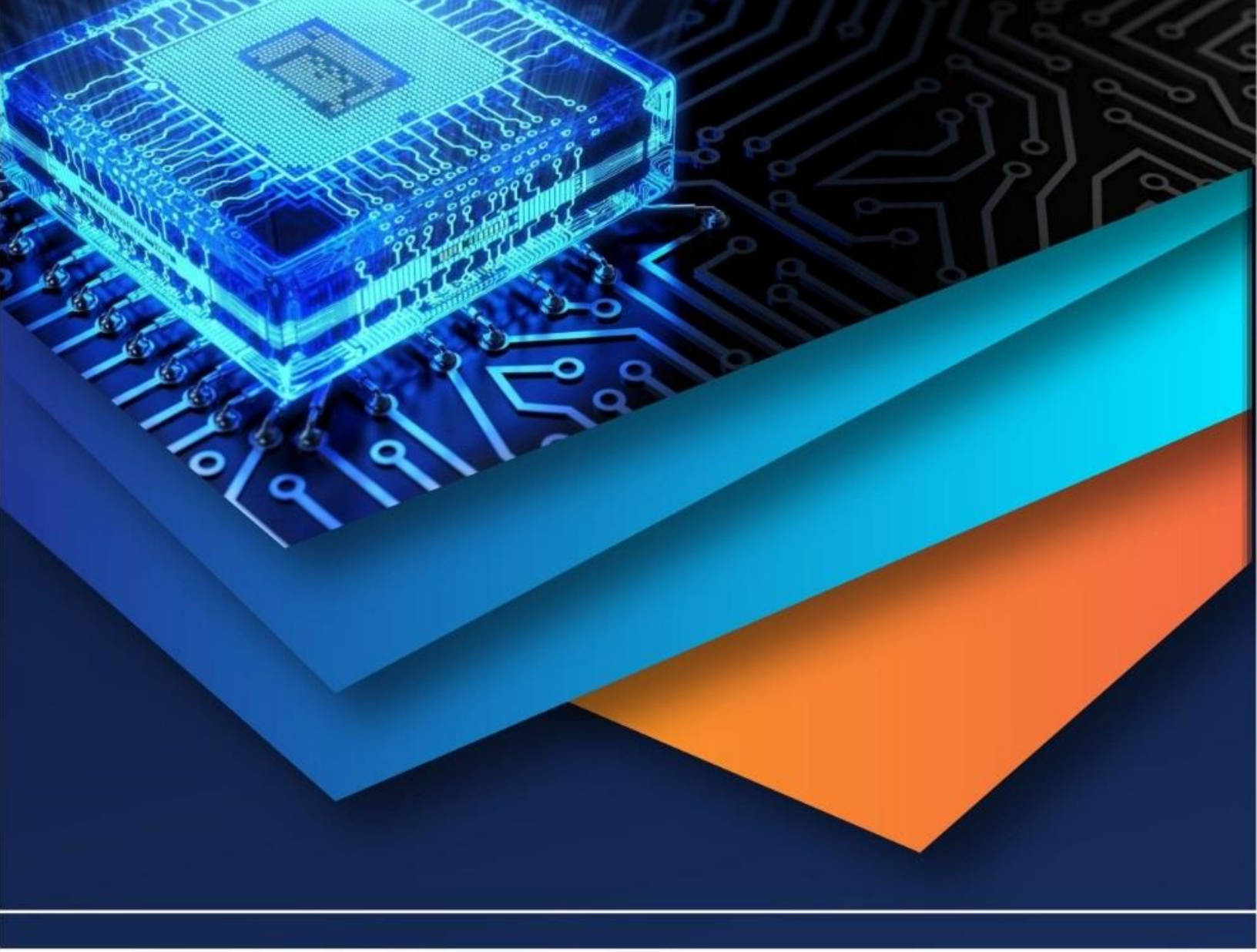
[7] Niphat Claypo Department of Computer Science, Saichon Jaiyen Department of Computer Science, "Automatic Keyword Selection for Sentiment Analysis using Class Dependency and Dissimilarity", 2016.

[8] Ghiassi, M. and Saidane, H., "A Dynamic Architecture for Artificial Neural Network", Neuro computing, (63), 2005.

[9] M. Ghiassi, M. Olschimke, B. Moon, P. Arnaudo, "Automated Text Classification using a Dynamic Artificial Neural Network Model", Expert Systems with Applications, (39), 2012.

[10] David Zimbra, M. Ghiassi, and Sean Lee, "Brand-Related Twitter Sentiment Analysis Using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks." 49th Hawaii International Conference on System Sciences (HICSS), IEEE, 2016.

[11] Shivani Pathak, Piyusha Mahajan, Ankita Patil, Rutuja Patil, MS.M.M.Swami, "Sentiment Analysis using Three Algorithm SVM, Naive Bayes,DAN2 "



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY