Sentiment Analysis of Twitter Datasets using different ML Techniques: A Systematic Study

Bipasha Hambir
AIML-CSE

West Bengal,India
hbipasha33@gmail.com

Abstract— There has been an exponential growth in the use of online resources, in particular social media and micro-blogging websites such as Twitter over the past decades. These resources offer a rich mine of marketing knowledge to organize. This project focuses on implementing a classifier using machine learning algorithms to extract sentiment of tweets. A major focus of this study was on comparing different machine learning algorithms based upon their performances. Also this approach allows to give a grade to the tweets based upon their intended sentiments which belong to one of the classes namely: negative, neutral, positive. From the evaluation of this study it can be concluded that the proposed machine learning techniques are effective and practical methods for sentiment analysis.

Keywords— Twitter, Sentiment analysis (SA), Machine learning, Support Vector Machine (SVM), Random Forest C.

*Introduction (*Heading 1*)*

Sentiment analysis is the process of recognizing and categorizing the sentiments represented in a text source. When analyzed, tweets are typically beneficial in providing a large volume of sentiment data. These statistics are valuable in determining public opinion on a variety of topics.

Opinion and sentimanetal mining is an important resarch areas because due to the huge number of daily posts on social networs, extracting people's opionin is a challenging task. About 90 percent of today's data has been provided during the last two years and getting insight into this large scale data is not trivial.

To compute the consumer perception, we must design an Automated Machine Learning Sentiment Analysis Model. It becomes challenging to develop models on them due to the existence of non-useful characters alongside helpful data. In this project, we will create a machine learning pipeline that uses many classifiers (Logistic Regression, Bernoulli Naive Bayes, SVM, RandomForest, KNN) as well as Term Frequency-Inverse Document Frequency to analyze the sentiment of tweets from the Sentiment140 dataset (TF-IDF). The accuracy and F1 Scores are then used to evaluate the performance of these classifiers.

In this paper, we will also discuss social network analysis and the importance of it, then we discuss Twitter as a rich resource for sentimental analysis. In the following sections, we show the high-level abstract of our implementation.

We will show some queries on different topics and show the polarity of tweets.

Data Description:

Twitter is an online news and social networking site where people communicate in short messages called tweets. These tweets can contain text, videos, photos or links. Tweets are short messages, restricted to 140 characters in length. Due to the nature of this microblogging service (quick and short messages), people use acronyms, make spelling mistakes, use emoticons and other characters that express special meanings. Following is a brief terminology associated with tweets. Emoticons: These are facial ex pressions pictorially represented using punctuation and letters; they express the user's mood. Target: Users of Twitter use the "@" symbol to refer to other users on the microblog. Referring to other users in this manner automatically alerts them. Hashtags: Users usually use hashtags to mark topics. This is primarily done to increase the visibility of their tweets.

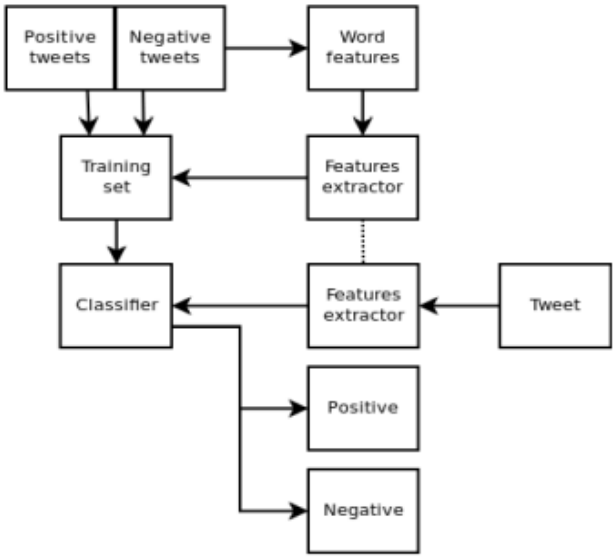
Data Preprocessing:

A tweet contains a lot of opinions about the data which are expressed in different ways by different users .The twitter dataset used in this survey work is already labeled into two classes viz. negative and positive polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features. The raw data having polarity is highly susceptible to inconsistency and redundancy. Preprocessing of tweet include following points,

- Remove all URLs (e.g. www.xyz.com), hash tags (e.g.#topic), emails)
- Correct the spellings; sequence of repeated characters is to be handeled
- Replace all the emoticons with their sentiment.
- Remove all punctuations , symbols, numbers
- Remove Stop Words
- Remove special and accented character.

TRAINING:

For tackling classification difficulties, supervised learning is an essential approach. Training the classifier makes it easier to anticipate unknown data in the future.



Classification:

Bernoulli Naive Bayes:

It is a probabilistic classifier that can learn the pattern of evaluating a group of previously classified texts [9]. It compares the contents of the documents to a list of terms to assign them to the appropriate category or class. Let d stand for the tweet, and c* for the class that d is assigned to. The count of feature (fi) is indicated by ni (d) in the preceding equation, and is present in d, which represents a tweet. The number m specifies the number of characteristics.

$$C^* = \arg mac_c P_{NB}(c \mid d)$$

$$P_{NB}(c \mid d) = \frac{(P(c))\sum_{i=1}^{m} p(f \mid c)^{n_{i(d)}}}{P(d)}$$

Maximum likelihood estimates are used to calculate the parameters P(c) and P(f|c), and smoothing is used to account for unobserved features. The Python NLTK package may be used to train and classify using the Nave Bayes Machine Learning approach.

Support Vector Machine (SVM):

A support vector machine examines data, defines decision boundaries, and performs computations in input space using kernels. The input data consists of two sets of m-dimensional vectors. Then each piece of data represented as a vector is assigned to a class. Following that, we discover a margin between the two classes that is unrelated to any document. The margin of the classifier is defined by the distance; boosting the margin lowers indecisive decisions. SVM also enables classification and regression, which are useful in statistical learning theory, and it aids in detecting the aspects that must be considered in order to properly comprehend it.

Logistic Regression:

The logistic function, which is at the heart of the procedure, is called logistic regression.

The logistic function, also known as the sigmoid function, was created by statisticians to characterize the features of population expansion in ecology, including rapid increase and reaching the environment's carrying capacity. It's an S-shaped curve that can transfer any real-valued integer to a value between 0 and 1, but never exactly between those two points.

1 / (1 + e-value),

where e is the natural logarithms' base (Euler's number or the EXP() function in your spreadsheet), and value is the numerical value you want to modify.

RandomForest Classifier:

Random forests, also known as random decision forests, is an ensemble learning method for classification, regression, and other problems that works by training a large number of decision trees. For classification tasks, the random forest's output is the class chosen by the majority of trees. The mean or average prediction of the individual trees is returned for regression tasks. Random decision forests address the problem of decision trees overfitting their training set. Random forests outperform decision trees in most cases, but they are less accurate than gradient enhanced trees. However, data features can influence how well they function.

The random forest training algorithm uses the common approach of bootstrap aggregating, or bagging, to train tree learners.

KNClassifier:

K Nearest Neighbor algorithm is a type of supervised learning technique that is used for classification and regression. It's a flexible approach that may also be used to fill in missing values and resample datasets. K Nearest Neighbor considers K Nearest Neighbors (Data points) to estimate the class or continuous value for the new Datapoint, as the name says.

The learning algorithm is:

1. Instance-based learning: Rather than learning weights from training data to predict output (as in model-based algorithms), full training instances are used to predict output for unknown data.

2. Lazy Learning: The model is not learned using training data before the prediction is required on the new instance, and the learning process is postponed until the prediction is asked.

3. Non-Parametric: In KNN, the mapping function has no specified form.

AdaBoost Classifier:

The AdaBoost algorithm, abbreviation for Adaptive Boosting, is a Boosting approach used in Machine Learning as an Ensemble Method. The weights are re-allocated to each instance, with higher weights applied to improperly identified instances. This is termed Adaptive Boosting. In

supervised learning, boost is used to reduce bias and variance. It is based on the notion of successive learning. Each subsequent learner, with the exception of the first, is grown from previously grown learners. In other words, weak students are transformed into strong students. With a little modification, the AdaBoost method works on the same idea as boosting. Let's take a closer look at this distinction.

XGBoost Classifier:
Extreme Gradient Boosting (XGBoost) is a distributed gradient-boosted decision tree (GBDT) machine learning toolkit that is scalable. It is the top machine learning package for regression, classification, and ranking tasks, and it includes parallel tree boosting.

Under the Gradient Boosting framework, XGBoost is an open-source software package that implements optimal distributed gradient boosting machine learning methods.

To understand XGBoost, you must first understand the machine learning ideas and methods on which it is based: supervised machine learning, decision trees, ensemble learning, and gradient boosting.

Supervised machine learning use algorithms to train a model to detect patterns in a dataset with labels and features, and then to predict labels on fresh dataset features using the learned model.

Prior Polarity Scaling:

A number of our features are dependent on the polarity of words before they are used. We draw inspiration from Agarwal et alwork's to determine the preceding polarity of words (2009). We utilise uWordNet to augment the Dictionary of Affect in Language (DAL) (Whissel, 1989). This dictionary of roughly 8000 English language terms offers a pleasantness value (R) to each word ranging from 1 (negative) to 3 (positive) (Positive). We begin by diving each score through the scale to normalise the results (which is equal to 3). Words with polarity less than 0.5 are considered negative, those with polarity more than 0.8 are considered positive, and the remainder are considered neutral. If a term cannot be located in the dictionary, we use Wordnet to find all synonyms.We then search DAL for each of the synonyms. If a synonym is identified in DAL, the original word is given the same pleasantness score as the synonym. If none of the synonyms appear in DAL, the term is unrelated to any previous polarity. We discovered antecedent polarity in 81.1 percent of the terms using the available data. Using WordNet, we can determine the polarity of the remaining 7.8% of terms.
As a result, about 88.9% of English language nouns have previous polarity.

Evaluation Of Sentiment Classification:

The performance of sentiment classification can be evaluated by using four indexes calculated as the following equations:

Accuracy = (TP+TN)/(TP+TN+FP+FN)
Precision = TP/(TP+FP)
Recall = TP/(TP+FN)
F1 = (2×Precision×Recall)/(Precision+Recall)

In which TP, FN, FP and TN refer respectively to the number of true positive instances, the number of false negativeinstances, the number of false positive instances and the number of true negative instances.

Result and Discussion:

We used the twitter dataset publicly made available by Kaggle. Analyses was done on this labeled datasets using various feature extraction technique. We used the framework where the preprocessor is applied to the sentences which make it more appropriate to understand. 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.

1.Support Vector Machine (SVM):

	precision	recall	F1-score
Train			
Test	0.85	0.85	0.85

2. Logistic Regression:

	precision	recall	F1-score
Train	0.92	0.95	0.93
Test	0.85	0.85	0.85

3. RandomForest Classifier:

	precision	recall	F1-score
Train	0.98	1.00	0.99
Test	0.84	0.84	0.84

4. BernoulliNB:

	precision	recall	F1-score
Train	0.86	0.93	0.90
Test	0.86	0.86	0.86

5.Decision Tree:

	precision	recall	F1-score
Train	0.94	0.98	0.96
Test	0.81	0.81	0.81

6.Comparison of accuracy of all models:

Model	precision	recall	F1-score
Dicision Tree	0.81	0.83	0.82
Logistic Regression	0.85	0.85	0.85
SVClassifier	0.85	0.85	0.85
XGBoost	0.84	0.82	0.82
AdaBoost	0.83	0.76	0.75
RandomForest	0.84	0.84	0.84
KNClassifier	0.76	0.73	0.72
BernoulliNB	0.86	0.86	0.86
LinearSVC	0.83	0.83	0.83

Challenges in Sentiment Analysis:

1. Recognizing subjective text parts: Subjective text parts provide sentimental information. In certain cases, the same term might be considered subjective, while in others, it can be considered objective. This makes it difficult to distinguish between objective and subjective text.

2. Stifled expressions: In certain phrases, just a portion of the content affects the document's overall polarity.

"This movie should be excellent," for example. It seems like a wonderful premise, with well-known stars and a good supporting cast.

In this scenario, a simple bag-of-words technique would label it good emotion, while the true attitude is negative.

3. Order Dependence: Sentiment Analysis/Opinion Mining requires Discourse Structure Analysis.

For example, A is better than B expresses the polar opposite of B is better than A.

4. Entity Recognition: Text regarding a certain entity must be separated and then sentiment toward it must be analysed.

"I despise Microsoft, but I enjoy Linux," for example.

It is labelled as neutral by a basic bag-of-words method, however it has a particular attitude for both entities involved in the sentence.

5. Create a classifier that distinguishes between subjective and objective tweets.

The majority of current research focuses on appropriately identifying positive and negative data. It's important to consider how to categorise tweets with and without sentiment.

Applications of Sentiment Analysis:

1.Applications that leverage website reviews:
On practically any topic, the Internet now provides a significant collection of reviews and feedback. This covers product reviews, political input, service complaints, and so on. As a result, a sentiment analysis system that can extract sentiments regarding a certain product or service is required. It will enable us to automate the process of providing feedback or ratings for a certain product, item, or service. This would meet the requirements of both users and vendors.

2. Domain across applications: Sentiment Analysis has aided recent studies in sociology and other sectors such as medicine and sports by revealing trends in human emotions, particularly on social media.

3. Smart Home Applications:
The technology of the future is intended to be smart homes.
People will be able to operate any element of their home with a tablet device in the future since complete homes will be networked.
There has been a lot of research recently on the Internet of Things (IoT). IoT would also benefit from sentiment analysis. For example, the home might change its atmosphere to provide a calming and peaceful environment based on the user's present attitude or emotion.
Trend prediction can also be done using sentiment analysis. Important data about sales patterns and consumer satisfaction can be gathered by tracking public views.

4. Sub-component Technology Applications:
In recommender systems, a sentiment predictor system can be useful. Items with a lot of negative comments or low ratings will not be recommended by the recommender system.
We come across abusive language and other bad features in internet conversation. These can be discovered simply by recognising a strong negative attitude and taking action to counteract it.

5.Business Intelligence Applications:
People nowadays have been observed to read reviews of things available on the internet before purchasing them. For many businesses, the success or failure of their product is determined by online opinion. As a result, sentiment analysis is crucial in business. Businesses also want to extract sentiment from online reviews to improve their products and, as a result, their reputation and customer satisfaction.

Conclusion:

We present a survey and comparison of existing strategies for opinion mining, including machine learning and lexicon-based approaches, as well as cross domain and cross-lingual methods and some assessment metrics, in this paper. Machine learning methods such as SVM and naive Bayes and Logistic Regression have the best accuracy and can be considered baseline learning methods, but lexicon-based approaches are very effective in some circumstances and require little effort in human-labeled documents, according to research findings. We also looked at how different features affected the classifier. We can deduce that the clearer the data, the more accurate the results. When compared to other models, the bigram model gives superior sentiment accuracy. We can deduce that the clearer the data, the more accurate the results. When compared to other models, the bigram model gives superior sentiment accuracy.

Reference:

[1] A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326 [2] R. Parikh and

M. Movassate, “Sentiment Analysis of User- GeneratedTwitter Updates using Various Classi_cation Techniques",CS224N Final Report, 2009

[2] (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2, 2019 361 | P a g e www.ijacsa.thesai.org A Study on Sentiment Analysis Techniques of Twitter Data Abdullah Alsaeedi1 , Mohammad Zubair Khan.

[3] S. R. Das and M. Y. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the web," Management science, vol. 53, no. 9, pp. 1375-1388, 2007.

[4] Sentiment Analysis of Twitter Data Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau.

[5]] Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior, 31, 527-541.

[6] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.

[7] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," Business horizons, vol. 53, no. 1, pp. 59-68, 2010.

[8] A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support vector machine," Teori dan Aplikasinya dalam Bioinformatika, Ilmu Komputer. com, Indonesia, 2003.

[9] Socher, Richard, et al. "Recursive deep models for semanticcompositionality over a sentiment Treebank." Proceedings of theConference on Empirical Methods in Natural Language Processing(EMNLP). 2013.

[10] Sentiment analysis on twitter, Avijit Pal, Argha Ghosh, Bivuti Kumar.