Comparative Study of Machine Learning Algorithm for Twitter Sentiment Analysis

Lokesh Pal
Galgotias University,
Greater Noida, U.P., India
pallokesh12@gmail.com

Ms. Rani Singh
Galgotias University,
Greater Noida, U.P., India

Md. Riyan Ansari
Galgotias University,
Greater Noida, U.P., India

Abstract- Over the past ten years, sentiment analysis—the automatic identification of positive or negative sentiment in text—has been a popular area of study. As online user activity is exploding with new technologies, people are more and more sharing their opinions about things on social media, review sites, and blogs. These opinions, both positive and negative, are stated about individuals, groups, locations, events, and ideas. The extraction of such sentiments from Twitter tweets is now possible through the use of machine learning and natural language processing (NLP) techniques, as well as other methods for managing of text. The difficulties of sentiment analysis, the methods that have been devised to address these difficulties, and our own approach to sentiment analysis on Twitter are all covered in this paper. Our method goes beyond simply categorizing thoughts as positive or negative; it also offers valuable information for forecasting, trend analysis, and product profiling. Initial findings suggest that this approach could be applied to social media sentiment research to meet corporate needs.

1. INTRODUCTION

Large volumes of reviews, opinions, and facts are being stored as raw data on e-service websites or social media every day. Proper styles were necessary in order to work with those raw data. Final styles can highlight adjectives, adverbs, verbs, or nouns. Research has recently found that adjectives and adverbs work better together than adjectives only in sentiment analysis (8). But no research has been done on every possible combination of verbs, adjectives, and adverbs. The theoretical study of some popular styles or offerings of sentiment analysis is discussed in this research. To incorporate new elements, the suggested approach weighs the benefits and drawbacks of the suggested designs. The new method arranges verbs, adverbs, and adjectives according to machine knowledge at document position. The preceding combinations that are considered for analysis are adverbs-verbs, adjective-adjectives-verbs, and adverbs-adjectives-verbs, and adverbs-adjectives. Conventional classifiers, such as Naive Bayes (NB), Decision Trees, and Linear Models, are employed for analysis and to prevent outcomes. But because to social media, people are more eager and willing than ever to share details about their lives, expertise, adventures, and research with the world. By sharing their thoughts and offering critiques on social gatherings, they actively participate in activities.

Businesses are encouraged to gather more information about their brands, products, and reputations through social media and public sharing of their abilities and interests. This allows them to get feedback and keep up their productive job. Sentiment analysis is obviously a key component of customer relationship marketing-focused companies and top career creative customer experience operations. also for businesses attempting to manage their personalities, find new prospects, and promote their products. Businesses wish to automate the processes of removing irrelevant content, understanding conversations, connecting the dots, and reciprocating. It is evident that many people are interested in sentiment analysis.

In our current day, sometimes referred to as the information age or knowledge society, access to large volumes of information is no longer an issue due to the astounding quantity of fresh content that is released online every day. For many businesses in this day and age, information has become the most valuable trading item. But if we can develop and use methods to find and gather pertinent data and information, then process it to turn it into knowledge with skill and agility, that's where we learn how this enormous quantity of information works precisely.

However, in most cases, these pertinent facts and figures are provided in plain English as unstructured text rather than in structured sources like tables or databases. The same sentiment can convey two different ideas in two different situations. Some people also use various dialect communications, shoptalk, and word abbreviations for convenience. Because sentiments are subjective, it is risky to analyze and categorize them in terms of their opposite, such as negative, positive, or neutral (2).

Maximum results at query time are determined by utilizing simple Boolean queries to express sentiments pertaining to a post on Twitter, Facebook wall post, etc. But this will not provide proper and current knowledge for aggregate sentiments and is not enough for solving the problems in sentiment analysis mentioned above. It must take precise care in solving the problems mentioned above to get precise information after sentiment analysis. Most other systems trying to solve these problems are still at the discussion level. Although some systems try to analyze sentiments of other languages, some solutions that solve some of the above limitations are also available in the commercial market.

This study proposes a process that can be used as a tool for sentiment research on social media on Twitter to address the problems outlined hereinabove. It also proposes a process by which a person can utilize people's stations to develop knowledge about their products and services that can be used in business settings.

2. LITERATURE REVIEW

Other related sentiment analysis studies are included in this chapter. Most of these techniques evaluate sentiments as either positive or negative, however some are at the research level and a couple are commercially available.

Adobe Social Analytics

Adobe Social Analytics assists companies in understanding the social media and it's effects by analysis of the manner in which social platforms and online communities shape marketing outcomes. By monitoring and studying these conversations, the maps social activity to the most important business metrics like revenue and brand value. It also monitors how companies interact with their customers online through social media—i.e. how Facebook updates can drive web traffic and inform purchasing behavior. To measure sentiment, Adobe employs natural language processing algorithms.[3]

Brandwatch Sentiment Analysis

One of the sentiment analysis program is Brandwatch, which was developed by a group of PhD students in the UK and is available for purchase. Determining whether sentiment can be neutral, negative, or positive is its primary purpose.[4]

Sentiment140

Sentiment140 is a web-based tool specifically designed to Twitter sentiment analysis. Designed by three computer Stanford University graduates in science, this tool aims on both English and Spanish tweets to check whether a brand, product, or subject is viewed positively, negatively, or objectively.[5]

TweetFeel

TweetFeel is an online tool that interprets sentiments in real-time based on Twitter. It gathers tweets about a particular word or subject and identify them as either positive or negative. TweetFeel uses machine learning–based sentiment analysis, giving a more representative picture of public opinion.

Gloss-Based Sentiment Classification

This approach to sentiment classification works better traditional content analysis and is focused on the opinions described in the text. It is based on a process that quantitatively translates the text "glosses" or dictionary definitions of words. These estimates are subsequently employed in a semi-supervised system for classifying the sentiment of each word.

Sentiment Analysis using Adjectives and Adverbs

While most sentiment analysis methodologies focus on adjectives, verbs, and nouns, this approach concentrates Adverb-Adjective Combinations (AACs) to evaluate the force of subjective statements. Instead of merely conferring for up sentiment scores of words, it proposes an axiomatic model varying as per how linguistically adverbs are classified. The system possesses three unique AAC scoring methods that meet these demands. What is special about this method is that it not only examines sentiment—it also employs the sentiment scores to produce product profiles, identify patterns, and make predictions for users.[12]

The revolutionary aspect of our system is that it accomplishes more than just to calculate sentiment; it also applies the sentiment scores to provide product profiles, trend forecasting, and user projections. What is unique about our system is that it is more advanced than elementary sentiment analysis. It creates detailed product profiles, analyzes trends, and provides customers with predictive insights based on sentiment scores.

3. APPROACH

The various methodologies of machine learning that can be used to conduct a sentiment analysis includes: Lexicon-based approaches which apply AFINN, SentiWordNet, or similar dictionaries with fixed sentiments. The methods calculate the sentiment of the body of text based on the sum of emotionally charged words. Although these methods are straightforward to implement and apply, they do not deal well with contextual nuance, sarcasm, or polysemy. One example illustrating this difficulty is: "I love waiting in long lines." In this case, "love," which at a glance seems like a positive sentiment, in reality is negative in context.

Machine Learning Approaches

Some algorithms of lexicon based sentiment analysis were further developed with the application of machine learning algorithms like Naive Bayes, Support Vector Machines (SVM) and Decision Trees. Such algorithms utilize features like word count and syntax.

In the case of Naive Bayes, sentiment analysis works well for classifying movie reviews as it uses the volume of words in a particular sentiment as the indicator for the classification system which is quite effective.

SVM is well known for its performance on high-dimensional text data and in managing noisy data, which is the case for most sentiment analysis, therefore it is mostly used for sentiment analysis tasks. Even with the gap in the approach's ability compared to the lexicon-based methods, the reliance on extensive unsupervised feature selection and rich annotated data sets is still ever-present.

Deep Learning Approaches

Sentiment analysis received an overwhelming benefit from automated, multi-layered sophisticated feature analysis made possible by deep learning.

A text like a paragraph may be considered a sequence of words and so, Recurrent Neural Networks (RNNs) are capable of handling such sequential information. As is the case with more complex AI, RNNs struggle with distant relationships. Sentiment analysis is better suited with Long Short Term Memory Units (LSTMs), which are more accommodating to information across longer sequences of memory.)[13]

Originally from image processing, Convolutional Neural Networks (CNNs) see to excel at text classification as well. This is the case with short pieces of text like tweets or reviews of items. These also form structures at a lower level and work towards figuring out what these structures mean. This phrase level identification consists of and is not limited to phrases that indicate some sort of feeling.[14]

Transformers and BERT are some of the most important transformations in the field of NLP. Operating on prior architectures, self-attention mechanisms is something that no one had thought was possible. Those meant that capturing quite

wide ranging dependencies were possible, more productively than earlier. Sentiment analysis, much like many other tasks has become significantly more efficient with the use of pretraining and finetuning BERT which has established new standards.[15]

3.1 Data Collection

The dataset contains positive and negative tweets collected from the NLTK library.[11] Both negative and positive tweets were sampled to create neutral messages.[8]

The data set is split into test and training sets for model evaluation and performance evaluation.

3.2 Preprocessing

Prior to analyzing the tweets, we used the following procedural steps:

- Removal of stock ticker, retweet label, link, and hashtag.
- Tokenization and lemmatization using the SpaCy library.[9]
- Removal of stopword and punctuation to retain significant words.

3.3 Feature Engineering

A frequency dictionary helped to monitor the word occurrence patterns in every sentiment class. Features were derived by tallying the frequency of words that correspond to positive, negative, and neutral sentiments, as well as a bias term.

3.4 Model Selection

Logistic regression was applied since it is easy and effective in solving multi-class as well as binary classification issues. The model applies a sigmoid activation function to estimate the probability of class membership for sentiment.

4. Design and Implementation

4.1 Preprocessing Pipeline

Preprocessing pipeline employs SpaCy for cleaning and tokenizing the text, followed by lemmatization and removing stopwords. All these steps normalize the text without its semantic meaning being changed.

4.2 Frequency Dictionary Construction

Word frequency counting was accomplished by reading the preprocessed tweets systematically and storing the occurrences in a dictionary of word pairs with sentiment labels.

4.3 Feature Extraction

Feature Extraction Each tweet can be represented as a vector with the following features:

- Constant component (constant 1).
- Positive word frequency.
- Negative word frequency.
- Neutral word frequency.

4.4 Logistic Regression

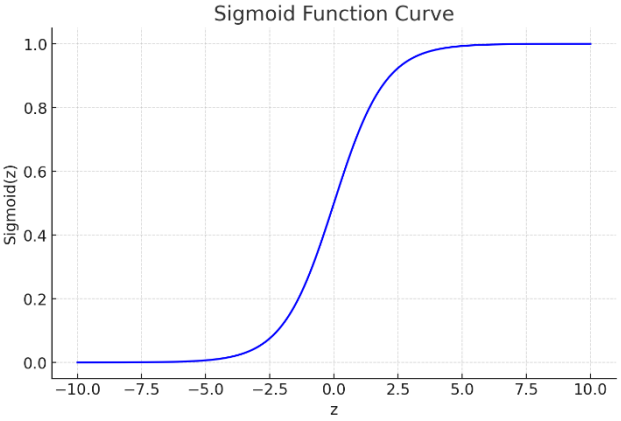
For both binary and multi-class classification tasks, statistical techniques like logistic regression are used.[10] Using a sigmoid function, it simulates the connection between input attributes and output labels.

The sigmoid function maps values to probabilities between 0 and 1, which are then used for classification based on predefined thresholds. For this project, positive sentiments were classified with probabilities above 0.6, neutral sentiments between 0.4 and 0.6, and negative sentiments below 0.4.

4.4.1 Key Concepts

1. Sigmoid Function formula:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



2. Decision Boundary:

- Classifies based on probability thresholds:
  - > 0.6 → Positive
  - < 0.4 → Negative
  - Between 0.4 and 0.6 → Neutral

3. Loss Function:

Measure prediction error.

$$J(\theta) = -\frac{1}{m} \sum [y \log(h) + (1 - y) \log(1 - h)]$$

4.5 Gradient Descent Optimization

The cost function, which quantifies the difference between predicted and actual values, is minimized using the iterative optimization technique known as gradient descent.

The cost function for logistic regression is defined as:

$$\text{Cost} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where is the predicted probability for sample . The gradient descent algorithm updates weights using the formula:

$$\theta := \theta - \eta \cdot \nabla L(\theta)$$

Where η is the learning rate.

In our implementation, we set the learning rate to 1e-9 and ran 1500 iterations to optimize the weights.

3.5.1 Working Principle

1. Initialize Parameters:
   - theta: Random weights is initialized to 0.
   - alpha: Rate of learning (step size).
   - num_iters: Number of iterations to update weights.
2. Compute Predictions (h):

$$h = \sigma(X \cdot \theta)$$

3. Update Weights (theta):
   Compute gradients based on the difference between predictions and actual labels

4.6 Prediction and Classification

Predictions were made by calculating the probability of each sentiment class using the sigmoid function. Thresholds were defined to classify tweets into positive (>0.6), negative (<0.4), and neutral (0.4–0.6) sentiments.

Probability Calculation:

- Using the trained weights:

$$z = X \cdot \theta$$

- Apply sigmoid:

$$p = \frac{1}{1 + e^{-z}}$$

5. Evaluation

5.1 Dataset Split

The dataset is then split into two parts namely training (80%) and testing (20%) subsets. Training data was used to optimize model parameters, while the test set was reserved for evaluation.

5.2 Performance Metrics

Model performance was assessed using the following metrics:

- Accuracy: Shows the percentage of tweets that are accurately classified.
- Precision: Assesses the ratio of true positives among predicted positives.
- Recall: Indicates the percentage of true positives to actual positives.
- F1-Score: Mean of precision and recall.

5.3 Results

Metric	Score
Accuracy	0.85
Precision	0.84
Recall	0.85
F1-Score	0.84

These results demonstrate the effectiveness of the model in classifying sentiments with high accuracy and balanced precision-recall performance.

6. Discussion

The model successfully leverages frequency-based features and logistic regression for sentiment classification. Key strengths include simplicity, interpretability, and fast training time. However, some limitations were identified:

- Sensitivity to vocabulary size and frequency imbalance.
- Difficulty in capturing complex linguistic patterns, such as sarcasm.
- Fixed thresholds for neutral sentiment, which may require dynamic adjustment.

Future work could explore deep learning approaches, such as recurrent neural networks (RNNs) or transformers, for improved accuracy in complex sentiment detection.

7. Conclusion

This paper demonstrates the feasibility of using logistic regression for multi-class sentiment analysis. By leveraging preprocessing techniques, frequency-based features, and gradient descent optimization, the model achieves competitive performance. The results highlight logistic regression as a strong threshold for sentiment analysis tasks, providing a foundation for further enhancements with advanced machine learning algorithms.

From the aspect of the project's top level, we gather information from social media platforms in order to extract sentiments from them and record those sentiments along with the details of the individuals who expressed them for future use.

8. Future Work

When implementing the sentiment module, we had to take into account a number of concerns, including the potential that user comments about a brand or product may not always be in English and may also contain emotional symbols or other languages, as well as the possibility that the comments may not accurately reflect the user's intended message. We had to take into account the fact that human language is not precise and clear, the uncertainty of the comment's wording, how a specific statement relates to previous comments, how to identify the entity, and the potential for people to use slang in their chats. Prior studies using R-Programming for Twitter data analysis (J.DavidSukeerthikumar et al.) provide alternative approaches that could complement our findings.[12]

Although there are still some problems with understanding natural language, the use of machine learning techniques could lead to more accurate findings if classifiers are built and trained on large labeled data sets.

Lastly, by combining user data with sentiment scores about a specific product or service, we can effectively profile the items, examine trends, and make predictions. Lastly, the system can tell you why a group of people of a certain occupation, location, and age perceive a certain commodity or service in a certain manner and how that's going to affect future information, which is of enormous value in business.

REFERENCES

[1] David Osimo and Francesco Mureddu, "Research challenge on Opinion Mining and Sentiment Analysis"

[2] Maura Conway, Lisa McInerney, Neil O'Hare, Alan F. Smeaton,Adam Berminghan, "Combining Social Network Analysis and Sentiment to Explore the Potential for Online Radicalisation," *Centre for Sensor Web Technologies and School of Law and Government*.

[3] Adobe® SocialAnalytics, powered by Omniture®.

[4] Brandwatch. [Online].http://www.brandwatch.com/

[5] Sentiment140.[Online]. http://www.sentiment140.com

[6] Fabrizio S. Andrea E., "Determining the Semantic Orientation ofTerms through," October 31– November 5 2005.

[7] Carmine C., Diego RFarah B., "Sentiment Analysis: Adjectivesand Adverbs are better," _ICWSM Boulder, CO USA_, 2006.
[8] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.
[9] SpaCy Documentation. (2023). Available at: https://spacy.io
[10] Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830.
[11] NLTK Documentation. (2023). Available at: https://www.nltk.org
[12] J.DavidSukeerthikumar, Y.Suteja, S.Supriya and A.Supriya, P.Uttejitha “Opinionanalysis on Twitter Data using R-Programming”.
[13] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
[14] Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746-1751.
[15] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.