Comparative Analysis of Sentiment Analysis based on tweets: Procedure, Various Approaches and its Challenges

Tiyashi Chatterjee, Atri Sanyal
NSHM Institute of Computing and Analytics, NSHM Knowledge Campus, Kolkata

Abstract

X Sentiment Analysis has become a pivotal area of research aimed at classifying tweets under categories based on the sentiments they seem to convey. It has often been used in understanding public opinion or user's response to various products, political elections, and predicting social and economic phenomenon such as stock market trends. The primary focus on X (previously known as Twitter) branches out from its vast and colossal user base, generating over 500 million tweets on a daily basis, providing a very rich and abundant source of data required for Sentiment Analysis. The primary approach involved or used towards Sentiment Analysis can be divided into a two-step process, firstly collecting tweets and then classifying its sentiment score otherwise known by polarity as positive, negative, or neutral. Various techniques such as Lexicon-Based Approaches and Machine Learning-Based Approaches have been used time and time again to achieve a more accurate and precise analysis on the tweets. Various challenges are also encountered when it comes to perform sentiment analysis on X (previously known as Twitter). One of the primary and demanding challenges of Sentiment Analysis is understanding the nuance behind a given piece of text. Other challenges such as word limit, emoticons continue to exist. Despite the progress that has been made in Sentiment Analysis of user responses and other text sources, sites such as Twitter present this unique challenge of understanding the nuance or the brevity of tweets. We have tried to compare the performance of some of the most widely used approaches for Sentiment Analysis based upon certain parameters. Finally, concluded that there is no clear winner among different approaches. Each approach has its own advantages and disadvantages and the choice depends on the type and uniqueness of the problem to be solved.

Keywords

Sentiment Analysis, Machine Learning, Lexicon-Based Approach, Twitter Data, Performance Metrics.

Introduction

Sentiment Analysis can be described as a field at the intersection of Natural Language Processing (NLP), Text Analysis, and Lexicon Understanding, playing an important role in identifying and providing sentiment scores to a piece of textual data. It involves the use of various algorithms, approaches and techniques aiming at determining whether a piece of text conveys positive, negative, or neutral emotions [2]. X (previously known as Twitter) a widely known, popular microblogging platform with a whopping 613 million user base, has become a focal point for Sentiment Analysis due to its user base, the number of tweets generated daily and the concise nature of tweets, being limited to 280 characters [7]. Twitter Sentiment Analysis aims to analyze the sentiments expressed in tweets to understand public opinion, consumer feedback, political views, and more. Using and combining machine learning algorithms and Lexicon-based approaches, Sentiment Analysis on Twitter can provide valuable insights or data for various domains. When saying that the analysis of sentiments on Twitter can help various domains, it can range from businesses in shaping their marketing strategies, monitoring brand perception, and gauging customer satisfaction to politics where Sentiment Analysis on X can help track political views. The challenges in Sentiment Analysis on Twitter stem from the brevity of tweets, cultural nuances, and the complexity of human language. The primary difficulty lies is understanding the emotion behind a given piece of text. Despite these challenges that continue to grow stronger with each passing day, sentiment analysis on X has proven to be a vital and useful tool for capturing the general mood of the online community. People have continuously been exploring innovative methods to enhance the performance of Sentiment Analysis on X, making it a dynamic, interesting, and evolving field in the realm of social media analytics [8].

Flow of Work while implementing Sentiment Analysis

No matter the approach is used for Sentiment Analysis, there exists a certain process that must be followed when building the model. The attached figure below will help us to get a primary idea of how one shall begin working.

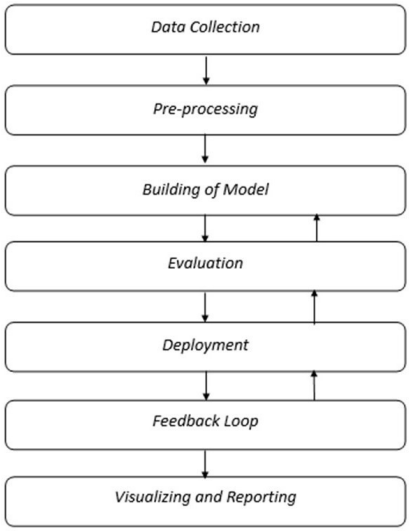

Figure 1: Flow of Work to Implement Sentiment Analysis

The above figure gives an estimated idea of the flow of work involved when building a Sentiment Analysis Model. Further, we will briefly see what happens at each step of the work flow.

1. Data Collection-Raw data is collected from X (previously known as Twitter). It can be done in primarily three ways-
   a. Using Twitter APIs such as 'Tweepy' and 'X API'
   b. By manually handpicking tweets from users and then creating a .csv or .xlsx file to store the tweets.
   c. Building a web scraper and storing data form a specific webpage by providing url.
2. Pre-Processing-Primarily involves getting rid of irrelevant, incomplete and duplicate data. It can be done in any of the following ways or all of them-
   a. Cleaning of Text-Involves removing mentions, hashtags, special characters etc. It involves handling of emoticons.
   b. Tokenization-It involves breaking down a piece of text into individual words or tokens.
3. Building of Model- Involves selecting one or combining more than one approach based on the requirements behind performing Sentiment Analysis. There are various approaches that can be used to build a Sentiment Analysis Model. Some of the most widely used approaches are-
   - Machine Learning Based
   - Lexicon Based
   - Ensemble Based

However, when conducting a comparative study, it has been seen that all approaches have their own sets of advantages and disadvantages, strengths and weaknesses [9]. Based on user requirements, any of the approaches are selected to build a Sentiment Analysis model.

It is important that we train the model for the data it may face when performing Sentiment Analysis on real life data. This involves feeding the data to the model, allowing it to find patterns and relationships between textual features and sentiments.

4. Evaluation- When building a Sentiment Analysis model, it is a crucial step that needs to be executed before the model is sent out in the real world to function. In this step, we primarily evaluate the said model by working with trial datasets. - Performance Metrics- Even though choosing the approach for Sentiment Analysis may differ for various requirements, it is pivotal that we compare the performance metrics to ensure that the model is best fit for the requirements. Some of the mostly used Performance Metrics used are-
   a. Accuracy
   b. Precision
   c. Recall
   d. F1-Score
6. Deployment- It is essential that once the model has undergone performance metrics such as Accuracy, Precision, Recall and received a satisfactory score, we deploy the model to see if it can handle real-life data as smoothly as seen in case of Evaluation and Deployment.
7. Feedback Loop- Once the model has undergone its deployment and evaluation, we can

or may need to train the model again by refining it further or make adjustments to achieve a better performance score and to help the model perform better. This loop is known as Feedback Loop[1].

Approaches and factors affecting Sentiment Analysis

While there are various approaches that exist for Sentiment Analysis, there are several factors that must be taken into consideration to receive better outcomes. Some of these factors are-

1. Data Quality and Quantity- Data is the very core and base for Sentiment Analysis. Hence, it is extremely crucial for the data that has been collected to be complete, accurate and relevant, as it can and may influence or alter the result of Sentiment Analysis. As important as the Data Quality is the Data Quantity. It is pivotal that we take more than enough data points so that the result of the Sentiment Analysis

model is not biased towards a particular sentiment polarity [6].

2. Data Cleaning- When collecting Data with the help of Twitter APIs or manually hand-picking them, there is a high possibility of finding irrelevant data, incomplete data, duplicate etc. If such data is not processed cleaned or removed properly, this data may affect the result and play a role in altering them. There are various techniques for Data Cleaning, some of the most widely used ones are
   a. Removing Duplicates
   b. Remove Irrelevant Data
   c. Standardize Capitalization
   d. Fix Structural Errors
   e. Delete Outliers, if any
3. Domain Specifications- When working with such voluminous amounts of data, it is extremely important that one is knowledgeable or aware of which domain they want to work with. Doing so can help filter the data as deemed necessary for the requirements. Not having a clear idea of the domain may require much more computational time and resources.
4. Ethics and Privacy-When working with data compiled or collected from social media, it is extremely crucial that we maintain, respect and give importance to user privacy as well as its other counterparts such as confidentiality throughout the entire process of Sentiment Analysis.
5. Evaluation of Model-There are various techniques that can be used to evaluate the performance of a model. Some of these are, Accuracy, Precision, Recall and F1-Score. It is crucial that the model's performance receives a satisfactory score when tested with these metrics, before being deployed.

Scope for Improvement- A Sentiment Analysis model may undergo changes and improvement, which may be iterative in nature as stated for the Feedback Loop, for the model to keep up with its performance, feedback etc.

A Comparative Study on the Various Approaches

Criterion	**Lexicon-Based Approach**	**Machine Learning-Based Approach**	**Ensemble-Based Approach**
Accuracy & Scalability	Might lack accuracy when handling complex language details.	Can achieve higher accuracy as they capture complex sentiment ex- pressions.	Outperforms individual models by combining predictions from multiple base models [4][5]
Interpretability	Offers high interpretability since sentiment scores are associated with words or phrases in the lexicon.	These models vary in interpretability. Ex: Decision trees are more interpretable compared to deep neural networks.	Ensemble methods sacrifice interpretability over accuracy.

Adaptation to Domain	They struggle with domain adaptation.	Can adapt to various domains with appropriate datasets and fine-tuning.	Can improve domain Adaptation by combining predictions.

Table 1: Comparison of different sentiment analysis approaches.

While all of the approaches have their own set of strengths and weaknesses, it is extremely crucial that the type of data provided and the requirements are well defined [10][11] Depending on the requirements one can choose to opt for one of the given approaches or a combination of them. However, it is crucial that the performance remains high.

Basic Sentiment Analysis Models used in Various Approaches

I. **Lexicon Based Approach**-This approach in Sentiment Analysis relies on the usage of a dictionary which contains either words or phrases annotated with sentiment scores. These sentiment scores define the word or phrase's polarity, indicating at the sentiment it carries which can either be positive, neutral or negative [3].
Advantages-

1. Legible in nature- Lexicon Based Approaches provide transparent and understandable results. The dictionary used contains sentiment scores that are readable and understandable to the human mind.
2. No Training Required- Lexicon Based Approaches do not require any training on labelled datasets, which can prove to be time consuming and may use up a lot of the resources and can also work with labelled and unlabeled datasets alike.
3. Medium Resistant to Noise- Lexicon Based Approaches can handle spelling errors and other grammatical mistakes as long as the sentiment determining phrases are available in the piece of text.

Disadvantages-

1. Lack of Robustness- Lexicon Based Approaches rely heavily on the available set of words or phrases available in the dictionary. Lack of coverage of words or phrases in the dictionary may cause the Sentiment Analysis system to provide less accurate results or even fail, making it less robust to handle all cases.
2. Scalability- Lexicon Based Approaches have proven to be efficient for small to moderate sized datasets, and may not perform as well with large or voluminous datasets.

II. **Machine Learning Based Approach**- This approach in Sentiment Analysis relies on training models to learn and recognize patterns and relationships in a given piece of text and further used to predict sentiment polarity i.e. positive, neutral or negative.

Advantages-

1. Flexible in Nature- Machine Learning Based Approaches have the ability to capture and understand more complex patterns in a given piece of text.
2. Accuracy- Since Machine Learning Based Approaches have been trained on large and a variety of datasets, they can achieve a much higher accuracy rate and have proven to be scalable.
3. Adaptive in Nature- Machine Learning Based Approaches tend to be adaptive in nature, which allows them to be re-trained with new data to be at par with changing patterns in text.

Disadvantages-

1. Problem of Overfitting- If Machine Learning Based Approaches catch noise or irrelevant data, it can overfit itself to the given data, affecting its performance.
2. Complexity in Calculation- Using Machine Learning Based Approaches with a vast number of features can take up a lot of time and resources.

III. **Ensemble Based Approach-** Ensemble Based Approaches rely on combining various in- dividual classifiers and models alike to improve accuracy and performance. Combining more than one model or classifier helps overcome individual shortcomings and hence promise a better performance altogether.

Advantages-

1. Higher Accuracy- Since Ensemble Based Approaches usually combine more than one classifier or model, it promotes a higher rate of accuracy.
2. Combining Predictions- Several prediction methods are combined to make the final sentiment predictions. Methods such as Voting, Stacking, Boosting etc. are used.
3. Reduction in number of predictions- Ensemble Based Approaches average or combines multiple individual predictions, resulting in taking into consideration all the predictions made by individual models.

Disadvantages-

1. Reduced Interpretability- Ensemble Based Approaches often forsake interpretation for increased accuracy rates.
2. Increase in Complexity- Combining two or more classifiers or models make the process of Sentiment Analysis using Ensemble Based Approaches much more complex.

References

1. Alsaeedi, A., & Khan, M. Z. (2019). A study on sentiment analysis techniques of twitter data. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *10* (2), 361. https://www.ijacsa.thesai.org
2. Berthelier, B. (2024). Social media sentiment analysis. *Encyclopedia*, *4* (4), 1590–1598. https://doi.org/10.3390/encyclopedia4040104

3. Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134. https://doi.org/https://doi.org/10.1016/j.knosys.2021.107134
4. Bordoloi, M., & Biswas, S. K. (2023). Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review*, *56*, 12505–12560. https://doi.org/10.1007/s10462-023-10442-2
5. Chalothom, T., & Ellman, J. (2015). Simple approaches of sentiment analysis via ensemble learning. In K. J. Kim (Ed.), Information science and applications (pp. 631–639). Springer Berlin Heidelberg.
6. Chaudhary, L., Girdhar, N., Sharma, D., Andreu-Perez, J., Doucet, A., & Renz, M. (2023). A review of deep learning models for twitter sentiment analysis: Challenges and op- portunities. IEEE Transactions on Computational Social Systems.
7. Kharde, V., Sonawane, P., et al. (2016). Sentiment analysis of twitter data: A survey of techniques. arXiv preprint arXiv:1601.06971.
8. Parveen, N., Chakrabarti, P., & Hung, B. T. e. a. (2023). Twitter sentiment analysis using hybrid gated attention recurrent network. *Journal of Big Data*, *10*, 50. https://doi. org/10.1186/s40537-023-00726-3
9. Tang, D., Qin, B., & Liu, T. (2015). Deep learning for sentiment analysis: Successful ap- proaches and future challenges. *WIREs Data Mining and Knowledge Discovery*, *5* (6), 292–303. https://doi.org/https://doi.org/10.1002/widm.1171
10. Xu, Q. A., Chang, V., & Jayne, C. (2022). A systematic review of social media-based senti- ment analysis: Emerging trends and challenges. *Decision Analytics Journal*, *3*, 100073. https://doi.org/https://doi.org/10.1016/j.dajour.2022.100073
11. Zhong, L., Li, L., & Zhong, X. (2024). An empirical study on sentiment intensity analysis via reading comprehension models. Proceedings of the 1st ACM Multimedia Workshop on Multi-Modal Misinformation Governance in the Era of Foundation Models, 23–28. https://doi.org/10.1145/3689090.3689390