Activity based

Project Report on

AI Business Intelligence

Project Phase - III

Submitted to Vishwakarma University, Pune

Under the Initiative of

Contemporary Curriculum, Pedagogy, and Practice (C2P2)

By

Shravan Sudhir Meshram

SRN No : 202101425

Roll No : 31

Div : E

Third Year Engineering

Faculty In charge:- Prof. Moumita Pal

Department of Computer Engineering

Faculty of Science and Technology

Academic Year

2023-2024 Term-II

Business Intelligence : Phase I

Project Name : Social Media Sentiment Analysis

Introduction:

Social Media Sentiment Analysis, as it provides a direct window into the thoughts and feelings of customers, stakeholders, and the general public. By integrating sentiment analysis into BI processes, organizations can enhance their decision-making processes, improve customer relations, and stay ahead of market trends. Social Media Sentiment Analysis within the BI framework involves the systematic extraction and interpretation of sentiments embedded in the vast sea of user-generated content across social media channels. This process empowers BI professionals to gauge public perception, track brand sentiment, and identify emerging trends with unprecedented granularity.

Social Media Sentiment Analysis is a pivotal component in the realm of Business Intelligence, offering a unique lens into the sentiments expressed across digital platforms. By employing advanced natural language processing and machine learning techniques, BI professionals can decipher the positive, negative, or neutral tones in user-generated content.

Problem Statement

Understanding customer sentiments on social media platforms is crucial for businesses to manage their online reputation, identify areas for improvement, and tailor marketing strategies. This project focuses on developing a Business Intelligence (BI) solution for Social Media Sentiment Analysis, extracting insights from social media data

---

Objective

The primary objective is to design and implement a BI system that integrates with social media data sources, performs sentiment analysis, and provides actionable insights into customer opinions, trends, and sentiment shifts.

Existing Work Done:

In this project, we undertake exploratory data analysis (EDA) on the Social Media Sentiment dataset. The process includes:

1. **Data Collection:** Collect social media data from various platforms, including customer reviews, comments, and mentions.

2. **Data Preprocessing:** Preprocess the social media data to handle noise, irrelevant information, and ensure data consistency. Clean text data, handle emotions, and address any data quality issues that may affect sentiment analysis.

3. **Feature Engineering:** Identify and engineer features that contribute to sentiment analysis. This may include sentiment scores, sentiment trends over time, and the identification of key topics or keywords associated with positive or negative sentiments.

4. **BI Dashboard Development**: Design and implement a user-friendly BI dashboard that visualizes key sentiment analysis metrics. Include components for monitoring overall sentiment trends, identifying sentiment influencers, and assessing the impact of marketing campaigns.

5. **Sentiment Score Calculation:** Implement features for calculating sentiment scores from social media text data. Utilize natural language processing (NLP) techniques to assess the polarity of customer opinions.

6. **Trend Analysis:** Develop features for trend analysis, tracking sentiment changes over time. Identify patterns, spikes, or dips in sentiment that may coincide with specific events, product launches, or marketing efforts.

7. **Influencer Identification:** Integrate features for identifying social media influencers who impact sentiment. This includes recognizing individuals or accounts whose opinions carry significant weight within the online community.

8. **Brand Mention Analysis:** Analyze brand mentions within the BI system. Understand the context and sentiment associated with brand mentions to assess the overall perception of the brand in the social media landscape.

Models Used in this Project:

SVM (Support Vector Machine) and Linear models for sentiment score analysis.

1. Linear Separability: SVM is particularly effective when the data is linearly separable, meaning the classes can be separated by a straight line (or plane in higher dimensions). In sentiment analysis, especially when using bag-of-words or TF-IDF representations, the feature space can often be linearly separable.
2. Robustness to Overfitting: SVM tends to generalize well to unseen data, especially when the margin between classes is maximized. This helps prevent overfitting, which is crucial in sentiment analysis where the model needs to perform well on new, unseen texts.
3. Kernel Trick: SVM can also handle non-linear decision boundaries through the kernel trick, where data is implicitly mapped into a higher-dimensional space where it can be linearly separated.

However, in sentiment analysis tasks, linear models can often perform well enough without the need for complex non-linear decision boundaries.

4. Interpretability: Linear models, such as logistic regression, are inherently interpretable. Each feature is assigned a weight, indicating its importance in predicting the sentiment score. This can be useful for understanding which words or features contribute most to positive or negative sentiment.
5. Efficiency: Linear models are computationally efficient, especially when compared to more complex models like neural networks. They can be trained quickly even on large datasets, making them suitable for sentiment analysis tasks where efficiency is important.
6. Scalability: Linear models are highly scalable, meaning they can handle large datasets with ease. This is important in sentiment analysis, where datasets can often be quite large due to the abundance of text data available.
7. Baseline Performance: Linear models often serve as a good baseline for sentiment analysis tasks. They are simple yet effective, and can often achieve reasonable performance without requiring extensive hyperparameter tuning or feature engineering.

Overall, SVM and linear models are popular choices for sentiment analysis due to their simplicity, efficiency, and effectiveness in many cases. However, it's always a good idea to experiment with different models and techniques to find the best approach for a specific dataset and task.

Exploring Other Models:

Some other popular models used in sentiment analysis:

1. Naive Bayes:

- Reasoning: Naive Bayes models are based on probabilistic principles and assume independence between features given the class label. They are simple and efficient, making them well-suited for sentiment analysis tasks, especially when dealing with text data. However, they may not capture complex relationships between features as effectively as SVM or linear models.

2. Random Forest:

- Reasoning: Random Forests are ensemble learning methods that combine multiple decision trees to improve performance and robustness. They can handle non-linear relationships and interactions between features well, which can be beneficial in sentiment analysis tasks with more complex data. However, they may be prone to overfitting on small datasets and can be computationally expensive compared to linear models.

3. Gradient Boosting Machines (GBM):

- Reasoning: GBM builds an ensemble of weak learners (typically decision trees) sequentially, each correcting the errors of its predecessor. GBM often provides better predictive performance compared to Random Forests, especially on structured data. However, they may require more hyperparameter tuning and can be more computationally intensive.

4. Recurrent Neural Networks (RNNs):

- Reasoning: RNNs, particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, are well-suited for sequential data like text due to their ability to capture contextual information and dependencies over time. They can potentially outperform traditional models like SVM and linear models by capturing more nuanced relationships within the text. However, they are more complex to train, require larger amounts of data, and can suffer from vanishing or exploding gradient problems.

5. Transformer-based Models (e.g., BERT, GPT):

- Reasoning: Transformer-based models have achieved state-of-the-art performance in various natural language processing tasks, including sentiment analysis. They can capture long-range dependencies

and contextual information effectively through self-attention mechanisms. These models often require significant computational resources and large amounts of labeled data for pre-training, but fine-tuning them on specific sentiment analysis tasks can yield excellent results.

In summary, while SVM and linear models offer simplicity, efficiency, and interpretability, other models like Naive Bayes, Random Forests, GBM, RNNs, and Transformer-based models may provide better performance by capturing more complex relationships and dependencies within the data. The choice of model depends on factors such as dataset size, computational resources, performance requirements, and interpretability needs. Experimentation with different models and techniques is often necessary to find the most suitable approach for a given sentiment analysis task.

Why Choosing SVM and Linear:

Choosing SVM and linear models for sentiment analysis can be justified based on several factors:

1. Interpretability: Linear models such as logistic regression provide straightforward interpretations of model coefficients. Each feature's weight indicates its importance in predicting sentiment, allowing analysts to understand which words or features contribute most to positive or negative sentiment. This interpretability can be valuable for gaining insights into the data and explaining model predictions.
2. Efficiency: SVM and linear models are computationally efficient, especially compared to more complex models like neural networks. They can be trained quickly even on large datasets, making them suitable for sentiment analysis tasks where efficiency is important. This efficiency allows for rapid experimentation and model iteration.
3. Baseline Performance: Linear models often serve as a good baseline for sentiment analysis tasks. They are simple yet effective, and can often achieve reasonable performance without requiring extensive hyperparameter tuning or feature engineering. Starting with a linear model provides a solid foundation for further experimentation with more complex models.
4. Robustness: SVM, in particular, is known for its ability to generalize well to unseen data and handle noise in the training data. By maximizing the margin between classes, SVM aims to find the decision boundary that best separates positive and negative sentiment instances. This robustness helps prevent overfitting and improves the model's performance on new, unseen texts.
5. Scalability: Linear models are highly scalable and can handle large datasets with ease. This scalability is important in sentiment analysis, where datasets can often be quite large due to the abundance of text data available. Linear models can efficiently process large volumes of data, making them suitable for real-world applications.
6. Ease of Implementation: Implementing linear models for sentiment analysis is relatively straightforward compared to more complex models like neural networks. Many libraries and frameworks provide built-in support for linear models, making it easy to train, evaluate, and deploy sentiment analysis models using SVM or linear regression.

```
Accuracy: 0.782312925170068

Classification Report:
              precision    recall  f1-score   support

    negative       0.83      0.74      0.78        39
     neutral       0.25      0.15      0.19        13
    positive       0.81      0.88      0.84        95

    accuracy                           0.78       147
   macro avg       0.63      0.59      0.61       147
weighted avg       0.76      0.78      0.77       147
```

```
Accuracy: 0.7891156462585034

Classification Report:
              precision    recall  f1-score   support

    negative       0.81      0.74      0.77        39
     neutral       0.40      0.31      0.35        13
    positive       0.82      0.87      0.85        95

    accuracy                           0.79       147
   macro avg       0.68      0.64      0.66       147
weighted avg       0.78      0.79      0.78       147
```