

Project Proposal and Plan:

Sentiment Analysis Combined with Predictive Modeling for Sales and Price Impact

1. Problem

In industries such as e-commerce, financial markets, and media, understanding the impact of public sentiment on sales or price trends is critical. For example, a sudden increase in positive customer reviews or social media mentions can boost product sales or affect stock prices, while negative sentiment may lead to a decline. However, many companies struggle to predict the extent to which these sentiments influence actual sales, stock price fluctuations, or consumer demand. This project aims to combine sentiment analysis with predictive modeling to forecast the impact of public sentiment on specific business metrics (e.g., product sales or price fluctuations), providing actionable insights for companies to respond strategically.

2. Proposed Method of Solution

This project will integrate sentiment analysis results from social media or customer reviews with predictive models to forecast sales trends or price fluctuations.

Key Steps:

- **Data Collection:** Collect sentiment-related data (e.g., social media posts, product reviews) and correlate it with relevant business metrics (e.g., product sales data, stock prices).
- **Sentiment Analysis:** Use **BERT** to classify posts or reviews into positive, neutral, or negative categories. This analysis will quantify how people feel about a product or event.
- **Predictive Modeling:**
 - **Independent Variable:** Sentiment scores derived from the text data (e.g., social media posts, customer reviews).
 - **Dependent Variable:** Business metrics such as sales volume, price trends, or demand.

These inputs will be fed into predictive models like **Linear Regression**, **Random Forest**, or **XGBoost** to forecast how changes in sentiment affect sales or prices over time.

- **Evaluation and Interpretation:** Once the predictive model is trained, we will interpret how sentiment changes affect the dependent variable, enabling companies to leverage this information in real time.
-

3. Potential Stakeholders

- **Product Sellers:** Can use sentiment-driven sales forecasts to anticipate demand and adjust pricing strategies.
 - **Financial Analysts:** Can predict stock price movements based on social media sentiment, enabling more informed trading decisions.
-

4. Potential Obstacles

- **Data Availability:** Aligning sentiment data (e.g., social media posts, reviews) with precise sales or pricing data may be challenging depending on the industry and source.
 - **Sentiment-Impact Variability:** Not all sentiment fluctuations will lead to proportional changes in sales or pricing, especially when external factors like market conditions or seasonal trends are at play.
 - **Model Interpretability:** While models like **Random Forest** and **XGBoost** are powerful, interpreting how much sentiment contributes to final predictions can be challenging. Feature importance metrics or **SHAP values** will be needed to break down the model's decisions.
-

5. Novelty of the Problem and Approach

While there has been extensive research on sentiment analysis and sales prediction, combining the two to predict **how sentiment affects sales or prices in real-time** is relatively novel. This project adds value by not only analyzing sentiment but also quantifying its impact on business outcomes, thus creating a more practical tool for decision-makers. By using a variety of machine learning models, we will provide more accurate and actionable insights into how real-time sentiment drives key business metrics.

Project Plan

1. Data Sources

- **Social Media Data:** Collect tweets or Reddit comments related to a specific product, company, or event via the **Twitter API** and **Reddit API**.
- **Product Reviews:** Extract customer reviews from platforms like **Amazon**, focusing on sentiment expressed about specific products.
- **Sales Data:** Collect publicly available sales data from e-commerce sites or collaborate with companies willing to provide historical sales metrics for their products.
- **Price/Stock Data:** Gather pricing trends from stock market platforms (e.g., **Yahoo Finance API**) or e-commerce platforms for specific products.

2. Dataset Creation

The dataset will include:

- **Timestamp:** Date and time of each social media post or review.
- **Sentiment Score:** The result of **BERT's** analysis (positive, neutral, or negative sentiment).
- **Business Metric:** Historical data for sales volume, stock price, or product demand, structured as a time series.
- **Additional Features:** External factors like seasonal trends, marketing events, or product releases to account for non-sentiment-based variations.

3. Data Organization

The data will be organized into a **time series dataset**, with each row representing a specific time period (e.g., daily or weekly), including sentiment scores and corresponding business metrics. Text data will be preprocessed by removing stop words, punctuation, and irrelevant symbols before feeding it into the sentiment analysis model.

4. Data Analysis Plan

- **Sentiment Analysis:** **BERT** will be used to classify the sentiment of each social media post or review. Sentiment scores will be aggregated by day or week for correlation with business metrics.
 - **Predictive Model:** A regression model, such as **Random Forest Regression** or **XGBoost**, will be trained to predict sales or price trends based on sentiment scores.
 - **Feature Importance:** An analysis will be conducted to measure how much of the variance in business metrics is explained by sentiment scores.
-

5. Model Type

- **BERT:** For accurate sentiment classification from textual data.
 - **Random Forest Regression** or **XGBoost:** To predict the impact of sentiment on business metrics like sales volume or price. These models handle complex relationships and are less prone to overfitting when tuned properly.
 - **Linear Regression:** As a baseline for comparison to more advanced machine learning models, providing an interpretable starting point.
-

6. Model Accuracy

- **Sentiment Model:** The **BERT** sentiment analysis will be evaluated using metrics like **accuracy**, **precision**, and **recall**, comparing it against labeled validation data.
 - **Predictive Model:** Accuracy of the sales/price prediction will be evaluated using metrics such as **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **R-squared (R^2)**. Feature importance metrics will also be used to assess the contribution of sentiment to overall predictions.
-

7. Existing Similar Models

While many studies focus on predicting sales or stock prices based on historical data alone, fewer have incorporated **sentiment as a key predictive feature**. By benchmarking our results against existing sales prediction models, we aim to improve upon them by incorporating real-time sentiment data.