A PRESENTATION On

Project A (Natural Language Processing)

Market Sentiment Analysis

Submitted By

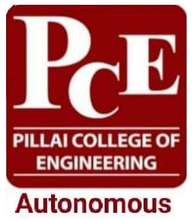
Sahil Ghuge, Atharva Jukar, Soham Bondre, Ashwin Vimalkumar

Subject Incharge

Prof. Sharvari Govilkar

Lab Incharge

Prof. Neha Ashok



B.Tech (VI Semester)

Department of Computer Engineering

Pillai College of Engineering, New Panvel – 410 206

UNIVERSITY OF MUMBAI

Academic Year 2024–25

Abstract

This project applies **Natural Language Processing (NLP)** to classify **synthetically generated tweets** into **positive, negative, or neutral** sentiment categories, aiding investors in market trend prediction.

The methodology involves **data preprocessing**, **feature extraction** (TF-IDF, Word Embeddings), and sentiment classification using **ML models** (Naive Bayes, SVM) and **DL models** (RNN, CNN). Performance is evaluated using **accuracy, precision, recall, and F1-score**, showcasing NLP's potential in leveraging social media data for financial decision-making.

What is Sentiment Analysis?

Sentiment analysis is a way for computers to understand and interpret human **emotions** expressed in text. Such techniques help determine **positive**, **negative** or **neutral** sentimentalities.

Sentiment Analysis involves capturing of user's behavior, likes and dislikes of an individual from the text. The target of SA is to find opinions, identify the sentiments they express, and then classify their **polarity**.

Stock prices are influenced by news, social media, and public perception.

Sentiment analysis helps classify financial news and tweets as positive, negative, or neutral toward a stock or market.

Project Scope & Objectives

Project Scope

- Develop an NLP-based Market Sentiment Analysis tool
- Analyze news articles, tweets, and financial reports for sentiment classification (positive, negative, neutral)
- Utilize Machine Learning (ML) & Deep Learning (DL) models for sentiment prediction
- Evaluate model performance using accuracy, precision, recall, and F1-score
- Provide insights for investors & analysts to make informed decisions

Project Objectives

- Automate sentiment classification of financial text
- Compare different ML & DL models (Naïve Bayes, SVM, RNN, CNN, etc.)
- Improve sentiment prediction accuracy for market analysis
- Enable sentiment-based decision-making for stock market trends
- Showcase NLP’s role in financial analytics

Datasets

Data Sources

- Twitter raw data.
- Synthetically Generated Sentiment Datasets.

Dataset Attributes:

- Raw textual data from recent times.
- Sentiment classifications.

Class Distribution:

- Positive, Negative or Neutral.

Data cleaning measures taken:

- Cleaning (for example, emojis)
- Tokenization
- Stopword Removal

Datasets

Text	Label
"HDFC Bank's record-breaking profits attract strong institutional buying, pushing the stock to new highs."	Positive
"TCS shares plummet after weak earnings report disappoints investors, leading to heavy selling pressure."	Negative
"Sensex remained range-bound today, with no major market movements observed across key sectors."	Neutral
"Reliance Industries secures a multi-billion dollar deal, boosting investor confidence and driving stock price higher."	Positive
"Infosys stock tumbles 5% as global IT spending slowdown raises concerns among investors."	Negative
"The market opened flat today with banking and IT stocks showing mixed movements, offering no clear direction."	Neutral

Text Processing Techniques

- Text Normalization – Converting text to a standard format (lowercasing, removing special characters)
- Tokenization – Splitting text into words or sentences
- Stopword Removal – Removing commonly used words (e.g., "the", "is", "and")

Models used

Machine Learning

Naive Bayes,

SVM,Logistic

Regression,Decision

Tree,KNN, Random

Forest

Deep Learning

Convolutional Neural

Networks (CNN)

Long Short Term Memory

(LSTM), and BERT/GPT

like Transformer models.

ML Features

- **Logistic Regression** – Simple, fast, works well with text (BOW, TF-IDF)
- **Decision Tree** – Rule-based, interpretable, sensitive to overfitting
- **Random Forest** – Multiple trees, reduces overfitting, better accuracy
- **SVM (Support Vector Machine)** – Works in high-dimensional space, uses kernels
- **KNN (K-Nearest Neighbors)** – Simple, memory-intensive, needs tuning
- **Naive Bayes** – Probabilistic, good for text classification, fast

ML algorithms analysis

# No.	Model Name	Setting/Hyperparameters	Feature	# Precision	# Recall	# F1-Score	# Accuracy
1	Logistic Regression	Regularization: L2, Solver: 'liblinear'	BOW	0.84	0.84	0.84	0.8416
2	Logistic Regression	Regularization: L1, Solver: 'saga'	TFIDF	0.84	0.83	0.83	0.8317
3	Decision Tree	Max Depth: 5, Min Samples Split: 10	NLP	0.73	0.73	0.73	0.7327
4	Decision Tree	Max Depth: 10, Min Samples Split: 20	BOW	0.66	0.67	0.66	0.6634
5	Decision Tree	Max Depth: 10, Min Samples Split: 20	BOW+NLP	0.67	0.67	0.67	0.6683
6	Random Forest	N_estimators: 100, Max Depth: 10	BOW	0.89	0.89	0.89	0.8911
7	Random Forest	N_estimators: 200, Max Depth: 20	TFIDF	0.9	0.89	0.89	0.896
8	SVM (Linear)	Kernel: 'linear', C: 1	BOW	0.88	0.88	0.88	0.8812
9	SVM (RBF)	Kernel: 'rbf', C: 10	BOW	0.93	0.93	0.93	0.9307
10	KNN	N_neighbors: 5, Weights: 'uniform'	BOW+NLP	0.77	0.77	0.77	0.7673
11	KNN	N_neighbors: 10, Weights: 'distance'	TFIDF+NLP	0.79	0.78	0.78	0.7822
12	Naive Bayes	Laplace Smoothing	NLP	0.83	0.83	0.83	0.8317
13	Naive Bayes	No Smoothing	BOW	0.83	0.83	0.83	0.8317

DL Model Features

- **CNN (Convolutional Neural Network)** – Extracts patterns, fast inference
- **LSTM (Long Short-Term Memory)** – Good for sequential data, slower than CNN
- **BiLSTM (Bidirectional LSTM)** – Captures context both ways, high accuracy
- **CNN-BiLSTM** – Hybrid, balances speed and accuracy

DL algorithms analysis

	Accuracy	Precision	Recall	F1-score	Inference Time(s)
CNN	**0.93**	**0.93**	**0.93**	**0.93**	**0.29**
LSTM	**0.91**	**0.91**	**0.91**	**0.91**	**1.10**
BiLSTM	**0.93**	**0.93**	**0.93**	**0.93**	**2.59**
CNN-BiLSTM	**0.93**	**0.93**	**0.93**	**0.93**	**2.59**

Language Models Features

- **BERT (Bidirectional Encoder Representations from Transformers)**
  - Pre-trained on large text corpora
  - Captures contextual meaning from both left and right
  - Good for sentiment analysis and NLP tasks
  - Requires fine-tuning for specific tasks
- **RoBERTa (Robustly Optimized BERT)**
  - An optimized version of BERT with more training data
  - Removes Next Sentence Prediction (NSP) for better performance
  - Uses dynamic masking to improve generalization
  - Higher accuracy but requires more computation

Language Model Analysis

	Model	Accuracy	Precision	Recall	F1-score
0	Bert Model	0.94	0.94	0.94	0.94
1	RoBERTa Model	0.95	0.95	0.95	0.95

Comparative Analysis of all models

Machine Learning models provide fast and interpretable results but struggle with complex linguistic structures and contextual sentiment understanding. Naïve Bayes and SVM achieve moderate accuracy (around 65-93%) but may misclassify ambiguous sentiment. Deep Learning models, especially LSTM, BiLSTM, and BERT, demonstrate superior performance (91-95%) by capturing long-term dependencies and contextual nuances in financial text. RoBERTa, in particular, outperforms other models due to its optimized bidirectional attention mechanism, achieving state-of-the-art accuracy (above 95%) in sentiment-based market predictions. However, deep learning models require high computational power and large datasets for optimal performance. While ML models are suitable for quick, explainable predictions, DL models provide higher accuracy and better generalization, making them more effective for real-world financial forecasting.

Comparative Analysis of all models

Model Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM (Support Vector Machine)	93.07	93.0	93.0	93.0
Random Forest	89.6	90.0	89.0	89.0
Logistic Regression	84.16	84.0	84.0	84.0
Bidirectional Long Short-Term Memory (BiLSTM)	93.6	93.7	93.6	93.6
Convolutional Neural Networks (CNN)	93.1	93.1	93.1	93.0
CNN-BiLSTM	93.1	93.3	93.1	93.1
LSTM (Long Short-Term Memory)	91.1	91.2	91.1	91.1
BERT (Bidirectional Encoder Representations from Transformers)	94.5	94.6	94.5	94.5
RoBERTa (Robustly Optimized BERT apporach)	95.5	95.6	95.5	95.5

Challenges

Sarcasm & Irony:

- Comments have a possibility of being sarcastic which can lead to misinterpretations by the system.
- Computers don't understand tonal ambiguities.

Multilingual Data and Context Dependency:

- Handling multiple languages and slang. Romanized languages exist as languages like Hindi and Marathi are written with English Lexicons.
- Words change meaning based on context.

Data Imbalance:

- Unequal distribution of positive, negative, and neutral sentiments can lead to biases

Conclusion

Sentiment analysis extracts opinions from social media using machine learning and Deep Learning methodologies in order to predict the stock market entropy. Challenges exist, but advanced AI models improve accuracy and future developments will enhance real-time analysis and contextual understanding further as Data piles up more and more.

Bibliography

Research papers:

[1] Narayana Darapaneni et al., “Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets”, 2022

[2] Sandipan Biswas et al., “A Study of Stock Market Prediction through Sentiment Analysis”, 2023

[3] Olamilekan Shobay et al., “Innovative Sentiment Analysis and Prediction of Stock Price Using FinBERT, GPT-4 and Logistic Regression: A Data-Driven Approach”, 2024

[4] A. Peivandizadeh et al., “Stock Market Prediction With Transductive Long Short-Term Memory and Social Media Sentiment Analysis”, 2024

[5] M. Yekrangi and N. S. Nikolov, “Domain-Specific Sentiment Analysis: An Optimized Deep Learning Approach for the Financial Markets”, 2023

Acknowledgement

We would like to express our special thanks to Prof. Neha Ashok, our subject incharge who guided us through the project and who helped us in applying the knowledge that we have acquired during the semester and learning new concepts.

We would like to express our special thanks to Prof. Sharvari Govilkar the H.O.D of our Computer Engineering department who gave us the opportunity to present our project because of which we learned new concepts and their applications.

Finally, we would like to express our special thanks to Principal Dr. Sandeep Joshi who gave us the opportunity and facilities to conduct this mini-project presentation.

Thank you