**Amrita Vishwa Vidyapeetham**

**Amrita School of Computing**

**Coimbatore**

**23CSE453 Natural Language Processing (2025-2026)**

**Case Study**

**Team No: 18**

|  |  |  |
| --- | --- | --- |
| Rollno | Name | Dept/Section |
| CB.EN.U4CCE22034 | Mangalasridharan Sankar Eswaran | CCE |
| CB.EN.U4CCE22053 | Vishal Thangakumar | CCE |

**Title of the Case Study:**

Sarcasm and Ambiguity-Aware Financial News Classification for Stock Market Impact Prediction

**Description of the complete work done** :

This project develops a Sarcasm and Ambiguity-Aware Financial News Classification System to predict stock market impact from news articles.

Using a subset (~1GB) of the FNSPID dataset, Financial PhraseBank, and Loughran-McDonald dictionary, the system processes 10,000 financial articles from 2018-2023.

Since FNSPID articles lacked labels, pseudo-labels (Positive/Neutral/Negative) were generated using FinBERT predictions on the summaries of the articles with confidence scores.

Both models were finetuned for the raw articles and found that FinBERT and Distil RoBERTa performed the same – 69% accuracy. This confirmed there should a lot of issues in the dataset, especially Ambiguities.

A custom rule-based ambiguity scoring system quantified syntactic, pragmatic, and lexical ambiguities, while T5-based sarcasm detection found negligible sarcasm in the dataset.

To resolve ambiguities, [AMBIGUOUS] special tokens were added near flagged sentences for pragmatic cases, and Named Entity Recognition (NER) addressed syntactic issues. DistilRoBERTa was selected for finetuning, because our motive was to improve prediction keeping the model size smaller. The model was fine-tuned on this ambiguity-handled dataset, achieving improved classification accuracy.

The research identifies a critical gap: lack of quantifiable ambiguity measures and properly labelled ambiguity datasets, suggesting that model improvements remain the primary path forward until better linguistic resources and ontologies are developed for financial text.

**Datasets used name with size:**

|  |  |  |
| --- | --- | --- |
| Dataset Name | Link(If Live data specify how much is free) | Size |
| FNSPID (Financial News and Stock Price Integration Dataset) | https://github.com/Zdong104/FNSPID\_Financial\_News\_Dataset | 10 GB |
| Financial PhraseBank | https://www.researchgate.net/publication/251231364\_FinancialPhraseBank-v10 | 5MB |

**Literature Survey paper title with author and journal details (minimum 6)**

1. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models – 2019 (arXiv.org)
2. FNSPID: A Comprehensive Financial News Dataset in Time Series-2024 (IEEE Transactions on Affective Computing.)
3. Comprehensive Review on Resolving Ambiguities in NLP- 2021 (AI Open - Sciencedirect)
4. Word Sense Disambiguation for Indic Language using Bi- LSTM (Advances in Intelligent Systems and Computing Springer)
5. To Word Senses and Beyond: Inducing Concepts with Contextualized Language Models (Published in EMNLP 2024 main conference proceedings.)
6. Analysis of Attention Mechanisms: The Case of Word Sense Disambiguation (Research Gate - Proceedings Conference on Machine Translation)

**Research Gaps in the allocated topic** :

1.Lack of Quantifiable Ambiguity Measurement Standards: There is no standardized, validated metric to measure different types of ambiguities (syntactic, semantic, pragmatic, lexical) in financial text.

2. Absence of Properly Labelled Ambiguity Datasets: Without true labelled data, you cannot definitively prove your ambiguity handling improves accuracy beyond what better pre-training or fine-tuning would achieve independently.

3. Insufficient Linguistic Resources for Financial Domain: NLP domain lacks domain-specific ontologies that map financial concepts, their relationships, and context-dependent meanings. Domain-specific ontologies that map financial concepts, their relationships, and context-dependent meanings

**Proposed Work with Novelty**:

1. First Attempt at Quantifying Multi-Dimensional Ambiguity in Financial News. We’re the first to systematically quantify three ambiguity types (syntactic, pragmatic, lexical) specifically for financial news classification.

Created a custom rule-based scoring system that assigns numerical ambiguity scores to financial articles,

1. Novel Hybrid Pipeline: Ambiguity Detection → Resolution → Classification

* Three-stage approach that's unique to your work:
  1. Detection: Rule-based ambiguity scoring (pragmatic/syntactic/lexical)
  2. Resolution: Different strategies per type:
     + Pragmatic: [AMBIGUOUS] special tokens near flagged sentences
     + Syntactic: NER-based entity disambiguation
  3. Classification: Fine-tuned FinBERT/DistilRoBERTa on ambiguity-enhanced data

1. Pseudo-Labeling Strategy for Unlabeled Financial News at Scale

* Generated pseudo-labels with FinBERT prediction on the article summaries along with confidence scores for 10K unlabeled FNSPID articles.
* Created a chunking strategy (512 tokens with variable stride overlap 128 -256) that preserves context across long articles help break long sentences into smaller chunks thus, helping model resolve syntactic and semantic ambiguities and meeting transformer input constraints.
* This enables training on large-scale recent financial news (2018-2023) that lacks manual annotations.

**Algorithm /Methods and Tools used**:

**Alg 1:** Dependency Parsing, Analyzes grammatical structure to identify syntactic relationships between words.

* Identifies sentence boundaries for proper tokenization
* Helps understand context around named entities
* Critical for handling complex financial terminology

**Alg 2:** Special Tokens [AMBIGUOUS] added to the flagged sentences after passing through our rule based pragmatic ambiguity quantifier.

**Alg 3:** NER (Named entity recognition), Identifies and classifies named entities in text.

Entity Categories:

* Organizations (ORG): Companies, agencies, institutions
* Persons (PERSON): People names
* Locations (GPE, LOC): Geopolitical entities, locations
* Temporal (DATE, TIME): Dates, times
* Numerical (MONEY, PERCENT, QUANTITY): Financial figures, percentages.

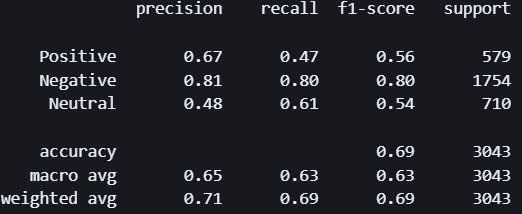
**Github Code Link :**

**Results :**

**Result Screen Shots : (With Description for each module and each case)**

\*Baseline models are not processed with NER and Special tokens whereas they are processed for Ambiguity handled models

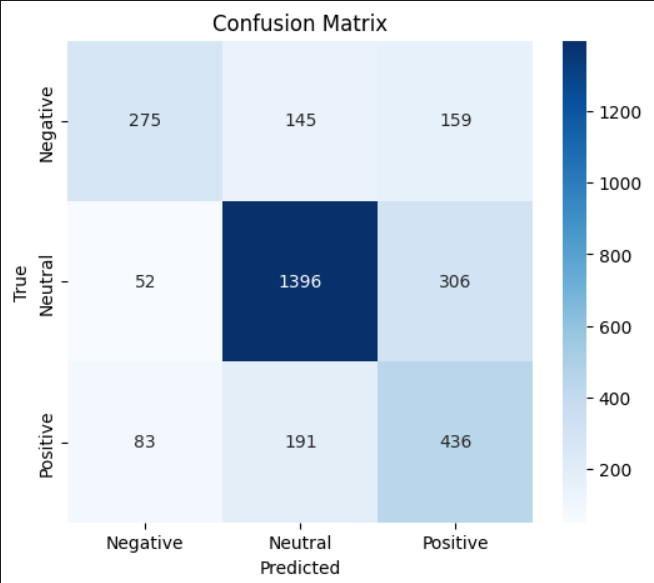
**CASE 1 (FinBERT Baseline Model):**

****

This shows FinBERT's performance without special tokens or NER preprocessing.

Performance by Class:

* Negative sentiment: Best performer (F1: 0.80, Precision: 0.81, Recall: 0.80) - Strong and balanced performance with 1754 test samples, indicating the model learned negative financial sentiment patterns well
* Positive sentiment: Poor performance (F1: 0.56, Precision: 0.67, Recall: 0.47) - Very low recall (47%) means the model misses over half of positive instances, likely confusing them with neutral
* Neutral sentiment: Weakest (F1: 0.54, Precision: 0.48, Recall: 0.61) - Low precision (48%) indicates many false positives; the model over-predicts neutral as a "default" class

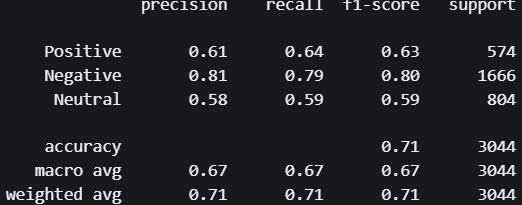
****

Correct Predictions (Diagonal):

* Negative: 1396/1754 correct (80% - strongest)
* Neutral: 436/710 correct (61% - moderate)
* Positive: 275/579 correct (47% - weakest)

Overall the Matrix suggests that baseline FinBERT without ambiguity resolution cannot distinguish subtle positive signals from neutral or even negative contexts.

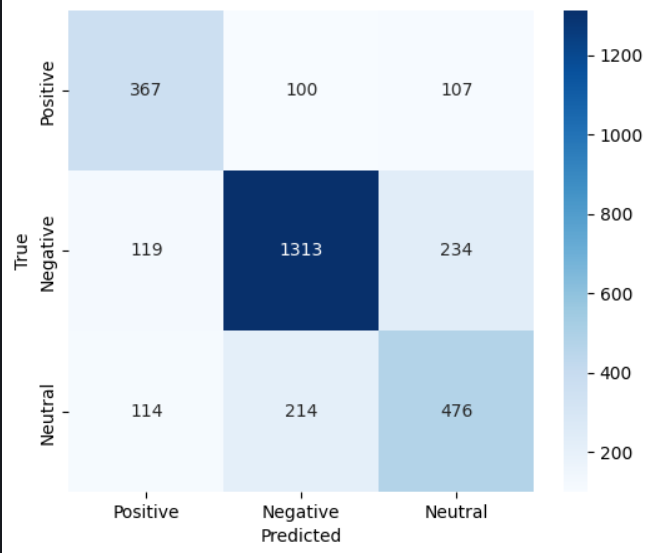
**CASE 2 (Roberta Baseline Model):**

****

This shows DistilRoBERTa's performance without special tokens or NER preprocessing (baseline model with 128-token stride)

Performance by Class:

* Negative sentiment: Strongest (F1: 0.80, Precision: 0.81, Recall: 0.79) - Consistent and balanced performance on 1666 samples, indicating robust negative pattern recognition
* Positive sentiment: Moderate (F1: 0.63, Precision: 0.61, Recall: 0.64) - Better recall than FinBERT baseline (64% vs 47%), but still struggles with precision
* Neutral sentiment: Weakest (F1: 0.59, Precision: 0.58, Recall: 0.59) - Low precision shows the model still defaults to neutral when uncertain, though less severely than FinBERT baseline

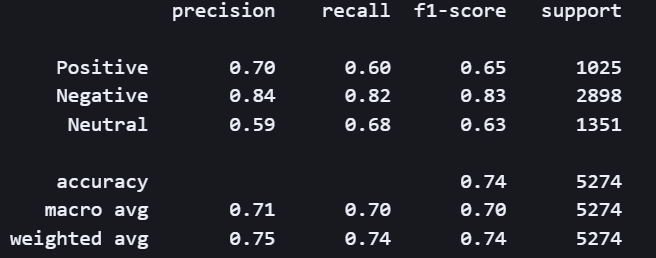
****

Correct Predictions (Diagonal):

* Negative: 1313/1666 correct (79% - strong)
* Positive: 367/574 correct (64% - moderate)
* Neutral: 476/804 correct (59% - weak)

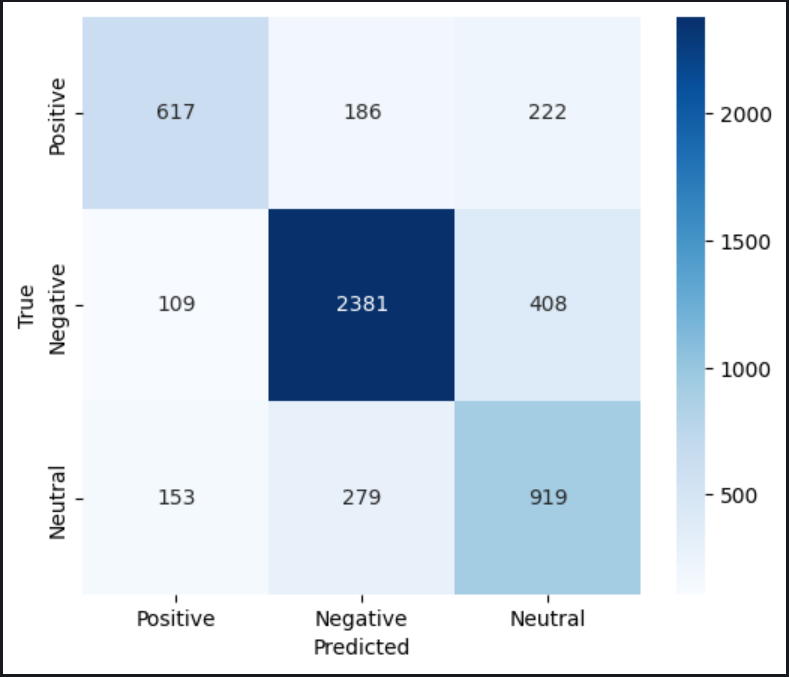
Positive detection improved: DistilRoBERTa correctly identifies 367/574 positives (64%) vs FinBERT's 275/579 (47%) - this is the most significant difference.

**CASE 3 (Distilled Roberta Ambiguity handled Model):**

****

This shows DistilRoBERTa's per-class performance on sentiment classification:

* Negative sentiment: Best performer (F1: 0.83, Precision: 0.84, Recall: 0.82) - The model excels at identifying negative financial news, likely because negative sentiment often uses distinctive keywords like "loss," "decline," "fall"
* Positive sentiment: Moderate performance (F1: 0.65, Precision: 0.70, Recall: 0.60) - Lower recall suggests the model misses positive instances, possibly confusing them with neutral
* Neutral sentiment: Weakest (F1: 0.63, Precision: 0.59, Recall: 0.68) - Low precision indicates many false positives, meaning the model over-predicts neutral when articles are actually positive/negative

****

Diagonal (correct predictions): Negative is strongest (2381 correct), Positive is weakest (617 correct).

Major confusion:

* 408 Negative articles misclassified as Neutral - suggests boundary ambiguity
* 279 Neutral and 186 positive misclassified as Negative - model has negative bias

Overall the Ambiguity handled model with NER and Special tokens applied shows improved accuracy and recall.

**Performance Metrics used:**

The model evaluation uses four key classification metrics:

Accuracy (74.27%)

* Measures overall correctness: proportion of correctly classified samples
* Formula: (TP + TN) / Total Predictions
* Suitable for balanced datasets; less reliable when class distribution is skewed

F1-Score (74.39% weighted)

* Harmonic mean of precision and recall
* Weighted average accounts for class imbalance by considering support (5number of samples per class)
* Formula: 2 × (Precision × Recall) / (Precision + Recall)
* Provides balanced measure between false positives and false negatives

Precision (74.81% weighted)

* Measures correctness of positive predictions
* Formula: TP / (TP + FP)
* Higher precision means fewer false positives (incorrect positive classifications)
* Class-specific: Negative sentiment shows highest precision (84%)

Recall (74.27% weighted)

* Measures ability to find all positive instances
* Formula: TP / (TP + FN)
* Higher recall means fewer false negatives (missed positive cases)
* Class-specific: Negative sentiment shows highest recall (82%)

**Performance measure result comparison Graph or table :**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Overall  Accuracy | Weighted F1 | Positive F1 | Negative F1 | Neutral F1 | Positive Recall | Negative Recall | Neutral Recall |
| DistilRoBERTa w/ Ambiguity Resolution | 74% | 0.74 | 0.65 | 0.83 | 0.63 | 60% | 82% | 68% |
| DistilRoBERTa Baseline | 71% | 0.71 | 0.63 | 0.80 | 0.59 | 64% | 79% | 59% |
| FinBERT Baseline | 69% | 0.69 | 0.56 | 0.80 | 0.54 | 47% | 80% | 61% |

**Future enhancements:**

1. Improved Data Utilization: Full FNSPID Dataset: Currently, only ~1 GB (10%) of FNSpid (~10 GB) was used. Scaling up to the entire dataset will improve coverage and generalization.
2. Additional Sources: Incorporate broader financial text sources (e.g., SEC filings, Bloomberg, Reuters) to reduce dataset bias.
3. Better Labeling Strategies: Move from pseudo-labeling with FinBERT to human-annotated or weakly supervised labels for higher-quality training data.
4. Long-Sequence Models: Replace standard transformers (512 tokens) with Longformer, BigBird, or LED (Longformer Encoder-Decoder), which can handle 4k–16k tokens efficiently.
5. Contextual Pragmatics: Integrate knowledge graphs or financial ontologies to improve pragmatic ambiguity detection.
6. Figurative Language Detection: Extend ambiguity detection to metaphors and idiomatic expressions, which are common in finance (e.g., “market rally,” “bear attack”).
7. Ensemble Models: Combine FinBERT, DistilRoBERTa, and domain-specific large models for better robustness.
8. Uncertainty-Aware Models: Instead of just confidence scores, use Bayesian neural nets or Monte Carlo dropout for calibrated confidence estimation.
9. Multilingual Extension: Extend sentiment and ambiguity analysis to multilingual financial texts (e.g., Chinese, German finance news).