People's Democratic Republic of Algeria
الجمهورية الجزائرية الديمقراطية الشعبية
Ministry of Higher Education and Scientific Research
وزارة التعليم والعالي البحث العلمي

Yahia Fares University of Medea
Faculty of Sciences
Department of Mathematics and Computer Science



Final Master Project

Submitted for Obtaining The Master's Degree in
Computer Science

Speciality: Applied Artificial Intelligence

Enhancing Financial Sentiment Classification via Fine Grained Emotion Modeling using ML, DL, and Attention Mechanism

Presented by:
Miss. Hanaa Aissani
Miss. Ibtihel Boudekhani

Proposed and Supervised by:
Dr. Ahmed Cherif Mazari
Dr. Nesrine Boudoukhani

2025/2026



Acknowledgements

At the end of this work, we would like to thank

***"Allah"** first, who gave us the will and patience to bring this work to completion.*

Elhamdulilah

We express our deepest gratitude and thanks to our supervisors ***Mr. Ahmed Cherif Mazari*** *and* ***Mrs. Nesrine Boudoukhani*** *for their invaluable guidance, exceptional supervision, and unwavering support throughout this research.*
Their expertise, patience, and dedication were instrumental in shaping the quality of this thesis.

We would like to thank the members of the jury for honoring us by evaluating our work — may they find here all expressions of consideration, gratitude, and distinguished respect.

We would also like to thank all our teachers who accompanied and guided us throughout our academic journey.

We acknowledge the use of Artificial Intelligence tools for writing assistance and language improvement.

Finally, we extend our profound gratitude to our family, as well as to all those who helped us, near or far, in completing this work.



Dedication

" وَآخِرُ دَعْوَاهُمْ أَنِ الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ "

All praise and gratitude belong to Allah, the Most Merciful, for granting me the strength, patience, and perseverance to complete this important milestone. Without His guidance and blessings, this achievement would not have been possible.

My deepest gratitude and heartfelt appreciation go to:

Mr. Ahmed Cherif Mazari, may Allah bless and reward you abundantly for your kindness and support. Throughout this journey, you were more than a supervisor; you were a mentor and a constant source of encouragement.

Mrs. Nesrine Boudoukhani, words cannot fully express my gratitude for your invaluable guidance, unwavering support, and genuine care. Thank you for always believing in me.

I dedicate this work, with immense love and gratitude:

To my beloved **mother**, the light of my life and the source of endless love and sacrifice. Your prayers were my greatest strength. May Allah bless you and grant you the highest rewards.

To my dear **father**, whose wisdom, sacrifices, and unconditional support illuminated my path. You have always been my role model, my pride, and my inspiration.

To my dear sister **Naziha**, my second mother, my guide, and a constant source of comfort and wisdom.

To my beloved brothers, **Abdelkarim and Rafik**, whose support and encouragement gave me strength during the most challenging moments.

To the little ones, **Yakin, Sirine, Arkam, and Ibrahim**, whose innocent smiles brought joy to my heart. May Allah bless you all.

To my beloved friend **Hanaa** who walked alongside me on this journey, sharing both difficulties and moments of success.

To my cherished **family**, my respected **professors**.

Finally, to every person who contributed to this achievement, directly or indirectly this achievement is also yours.

Ibtihel ♡



Dedication

First and foremost, I praise Allah Almighty and thank Him for granting me the success, patience, and determination that helped me complete this humble work.

I dedicate this work to:

To my honorable supervisor, Professor **Ahmed Cherif Mazari** Who was not merely a supervisor fulfilling his duty, but a wise guide and a true supporter. His great merit and immense effort were instrumental in completing this humble work, and no words could ever fully express the depth of my gratitude and appreciation for him. May Allah protect him and reward him with the best of rewards.

To the co-supervisor, Professor **Boudoukhani Nesrine**, Who enriched and refined this work through her valuable guidance and precise observations.

To the esteemed laboratory supervisor, **Khadidja Louz**, Who generously opened the doors of her office to us and placed her trust in us, this was an immense source of support in the completion of this work.

To my beloved **mother**, No matter how many words I am given, they will never be enough to thank you as you deserve. Your love surrounds me from every side, your tenderness lights my way, and you have always been the source of my strength and my refuge in moments of weakness.

To my beloved **father**, You are my first pillar and my unshakeable foundation. Through your boundless support, sincere love, deep faith in me, and the sacrifices you never hesitated to make in clearing every obstacle from my academic path, I have become who I am today.

To my dear sisters **Selma** and **Amel**, my beloved **grandmother**, and my dear aunt **Houria** — For your unwavering moral support, your precious belief in me, your wholehearted love, and your beautiful patience.

To all my **friends**, For your kind companionship and constant encouragement, and in memory of our shared laughter and the beautiful days we spent together.

To my dear **professors**, For the beneficial knowledge you imparted and the wise advice that illuminated my path.

To my dear husband **Chouaib**, my dear **family**, and to all who contributed to the completion of this thesis, whether directly or indirectly.

And finally, to my partner on this journey, **Ibtihel** — My companion with whom we shared a single road, full of hardship and stumbles. We were each other's support in every moment of weakness, and whenever one of us faltered, the other reached out her hand to lift her up. We struggled together, we persevered together, until we finally reached what we had always dreamed of. To you, my deepest love and most sincere appreciation.

To all those I love, and all those who love me.

Hanaa ♡

Abstract

Financial sentiment analysis is essential for investment decisions and market prediction. However, traditional methods that classify text into positive, negative, or neutral often fail to capture nuanced emotions that drive investor behavior.

To address these limitations, this thesis introduces an innovative hybrid framework that incorporates fine-grained emotion modeling to enhance financial sentiment classification. The core methodology is structured around three complementary approaches, beginning with Approach 1 (Traditional Financial Sentiment Representations), which utilizes FinBERT through both feature extraction and fine-tuning strategies, alongside general-domain models(RoBERTa, XLM-RoBERTa, and DeBERTa), to capture deep contextual financial semantics. This is complemented by Approach 2 (Emotion-Based Representations), which extracts fine-grained emotion vectors using a RoBERTa model pre-trained on the GoEmotions dataset to model underlying textual affect. Finally, Approach 3 (Hybrid Feature Fusion Strategies) integrates the deep contextual representations from the financial models with the fine-grained emotion vectors.

The proposed framework was evaluated across three benchmark financial datasets: Financial PhraseBank (100% and 50%), Twitter Financial News, and the Gold dataset. Classification tasks were executed using traditional Machine Learning algorithms, Deep Learning architectures, and attention mechanisms, followed by a comprehensive comparative analysis between the three developed approaches. Results demonstrated that incorporating fine-grained emotional dimensions substantially enhanced classification accuracy and robustness, bridging the gap between financial textual semantics and investor sentiment dynamics, a benefit most clearly reflected within Approach 3, where Fine-Tuning (FT) consistently surpassed Feature Extraction (FE) across all backbone models. The fusion of RoBERTa with RoBERTa-GoEmotion embeddings delivered the strongest results on PhraseBank 100% (LightGBM, F1-W: 99.55%), PhraseBank 50% (XGBoost, F1-W: 92.85%), Gold (BiLSTM, F1-W: 96.59%), and Twitter (LightGBM, F1-W: 92.92%), establishing RoBERTa as the most consistently effective backbone for emotion-augmented financial sentiment classification.

Keywords: Financial Sentiment Analysis, Fine-Grained Emotion Modeling, FinBERT, RoBERTa, XLM-RoBERTa, DeBERTa, GoEmotions, Feature Fusion, Machine Learning, Deep Learning, Attention Mechanism .

ملخص

تحليل المشاعر المالية أمر أساسي لاتخاذ قرارات الاستثمار والتنبؤ بحركة الأسواق. ومع ذلك، فإن الطرق التقليدية التي تصنف النصوص إلى إيجابية أو سلبية أو محايدة غالبًا ما تفشل في التقاط العواطف الدقيقة التي تؤثر على سلوك المستثمرين.

لمعالجة هذه القيود، تقدم هذه المذكرة إطارًا هجينًا مبتكرًا يدمج نمذجة العواطف الدقيقة لتعزيز تصنيف المشاعر المالية. وتتمحور المنهجية الأساسية حول ثلاثة مناهج متكاملة؛ تبدأ بـ المنهج الأول (تمثيلات المشاعر المالية التقليدية)، والتي تعتمد على نموذج FinBERT من خلال استراتيجيتي استخراج الميزات والضبط الدقيق، إلى جانب نماذج أخرى عامة المجال (DeBERTa, XLM-RoBERTa و RoBERTa)،لالتقاط الدلالات السياقية العميقة للمجال المالي. تليها المنهج الثاني (التمثيلات القائمة على العواطف)، والتي تستخرج تمثيلات عاطفية تعتمد على نمذجة العواطف الدقيقة باستخدام نموذج RoBERTa المدرب مسبقًا على مجموعة بيانات GoEmotions لنمذجة التأثير العاطفي الكامن في النصوص. وأخيرًا، يدمج المنهج الثالث (استراتيجيات دمج الميزات الهجينة) التمثيلات السياقية العميقة المستمدة من النماذج المالية مع التمثيلات العاطفية.

تم تقييم الإطار المقترح عبر ثلاث مجموعات بيانات مالية مرجعية (50% , 100%) Financial PhraseBank Dataset , Twitter Financial News Dataset و Gold Dataset وقد نُفِّذت مهام التصنيف باستخدام خوارزميات التعلم الآلي التقليدية، وبنيات التعلم العميق، وآليات الانتباه، تلاها تحليل مقارن شامل بين المناهج الثلاثة المطورة.أظهرت النتائج أن دمج نمذجة العواطف الدقيقة حسّن بشكل كبير من دقة التصنيف ومتانته، مما ساهم بفعالية في سد الفجوة بين الدلالات النصية المالية وديناميكيات مشاعر المستثمرين. وقد ظهر هذا التحسن بوضوح أكبر في المنهج الثالث، حيث تفوق الضبط الدقيق باستمرار على استخراج الميزات عبر جميع النماذج الأساسية.

وحقق دمج RoBERTa مع تضمينات RoBERTa-GoEmotion أفضل النتائج على 100% Financial PhraseBank (LightGBM، F1-W: 99.55 %)، و 50% Financial PhraseBank (XGBoost، F1-W:92.85%)، و Gold dataset (BiLSTM، F1-W: 96.59 %)، و Twitter Financial News (LightGBM، F1-W:92.92%).

وهذا يؤكد أن RoBERTa هو النموذج الأساسي الأكثر فعالية وثباتًا في تصنيف المشاعر المالية المعزز بالعواطف.

الكلمات المفتاحية: تحليل المشاعر المالية، نمذجة العواطف الدقيقة، FinBERT، RoBERTa، GoEmotions، RoBERTa, XLM-RoBERTa, DeBERTa, دمج الميزات، التعلم الآلي، التعلم العميق، آليات الانتباه.

Contents

List of Figures x

List of Tables xi

List of Acronyms xii

General Introduction 1

1 Background and Literature Review 4
1.1 Introduction . . . 4
Part 1: Background of AI . . . 4
1.2 Machine learning . . . 4
1.2.1 Support Vector Machine (SVM) . . . 5
1.2.2 Random Forest (RF) . . . 5
1.2.3 Logistic Regression (LR) . . . 5
1.2.4 eXtreme Gradient Boosting (XGBoost) . . . 5
1.2.5 Categorical Boosting Gradient (CatBoost) . . . 6
1.2.6 Light Gradient Boosting (LightGBM) . . . 6
1.3 Deep Learning . . . 6
1.3.1 Convolutional Neural Networks (CNN) . . . 6
1.3.2 Recurrent Neural Networks (RNN) . . . 7
1.3.3 Long Short-Term Memory (LSTM) . . . 7
1.3.4 Bidirectional LSTM (BiLSTM) . . . 7
1.3.5 Gated Recurrent Unit (GRU) . . . 8
1.3.6 Bidirectional GRU (BiGRU) . . . 8
1.4 Word Embeddings to Contextual Models . . . 8
1.5 Attention Mechanisms . . . 9
1.5.1 Self-attention . . . 9
1.5.2 Multi-Head Attention . . . 10
1.5.3 Hierarchical Attention . . . 10
1.6 Transformer Architecture . . . 10
1.6.1 Transfer Learning in NLP . . . 11
1.6.2 Pre-trained Language Models . . . 11
1.6.3 Domain-Specific Language Models . . . 12
Part 2: Financial Sentiment Analysis and Emotion Modeling Foundations . . . 12
1.7 Financial Sentiment Analysis (FSA) . . . 12
1.7.1 Definition . . . 12
1.7.2 Specificities of the Financial Domain . . . 12
1.7.3 Objectives of FSA . . . 13
1.7.4 Key Challenges . . . 13
1.8 Landscape of Financial Sentiment Datasets . . . 13
1.8.1 Benchmark Research Datasets . . . 13
1.8.2 Social Media Datasets . . . 14

1.8.3 Sentiment Analysis in the Gold Commodity Market . . . 14
1.8.4 Synthesis of Data Resources . . . 15
1.9 Fine-Grained Emotion Modeling . . . 15
1.9.1 Comparison between Sentiment Analysis and Emotion Polarity . . . 15
1.9.2 GoEmotions Taxonomy . . . 15
1.10 Related Work . . . 16
1.10.1 Financial Sentiment Analysis . . . 16
1.10.2 Fine Grained Emotion Recognition . . . 16
1.10.3 Literature Gap . . . 18
1.11 Conclusion . . . 18

2 Research Methodology and Proposed Approaches 20
2.1 Introduction . . . 20
2.2 Problem statement . . . 20
2.3 Proposed Framework Overview . . . 21
2.4 GoEmotions dataset . . . 22
2.5 Financial Sentiment Datasets . . . 23
2.5.1 Financial PhraseBank (50%, 100% agreement) . . . 23
2.5.2 Financial Sentiment Analysis GOLD Headlines . . . 24
2.5.3 Twitter Financial News . . . 24
2.5.4 Distribution of classes in the different datasets . . . 25
2.6 Proposed Approaches . . . 25
2.6.1 Approach 1: Traditional sentiment analysis . . . 25
2.6.2 Approach 2: Fine Grained based sentiment analysis . . . 26
2.6.3 Approach 3: Hybrid Feature Fusion . . . 27
2.7 Classification setup . . . 28
2.8 Evaluation Metrics . . . 29
2.8.1 Confusion matrix . . . 29
2.8.2 Accuracy . . . 30
2.8.3 F1-score Macro . . . 30
2.8.4 F1-score Weighted . . . 30
2.8.5 Matthews Correlation Coefficient (MCC) . . . 31
2.9 conclusion . . . 31

3 Experiments and Results 32
3.1 Introduction . . . 32
3.2 Development Environment . . . 32
3.2.1 Hardware Specifications . . . 32
3.2.2 Software Environment & Libraries . . . 32
3.3 Experimental Setup . . . 33
3.3.1 Data splitting . . . 33
3.3.2 Hyperparameters . . . 34
3.4 Experiment Results of Approach 1 . . . 36
3.4.1 Feature Extraction . . . 36
3.4.2 Fine-Tuning . . . 38
3.4.3 Discussion Results of Approach 1 . . . 40
3.5 Experiment Results of Approach 2 . . . 41
3.5.1 Discussion Results of Approach 2 . . . 44
3.6 Experiment results of Approach 3 . . . 44
3.6.1 Concatenation FinBERT Features Extraction with Roberta - GoEmotion (27-768) . . . 44
3.6.2 Concatenation Fine Tuning FinBERT with Roberta - GoEmotion (27-768) 47

3.6.3 Concatenation Fine Tuning RoBERTa with Roberta - GoEmotion (27-768) . . . 50
3.6.4 Concatenation Fine Tuning DeBERTa with Roberta - GoEmotion (27-768) . . . 52
3.6.5 Concatenation Fine Tuning XLM-RoBERTa with Roberta - GoEmotion (27-768) . . . 55
3.6.6 Discussion Results of Approach 3 . . . 57
3.7 Global Study Discussion . . . 58
3.7.1 Emotion-Sentiment Interaction Analysis . . . 58
3.7.2 Comparison Between Proposed Approach . . . 64
3.7.3 Error Analysis . . . 65
3.8 Comparative With Literature Review . . . 68
3.9 Conclusion . . . 69

General Conclusion **70**

Bibliography **74**

List of Figures

1.1 Architecture of a Random Forest Model (Khan et al., 2021) . . . 5
1.2 Architecture of a CNN-1D Model (Hoa et al., 2024) . . . 7
1.3 Architecture of a LSTM Model (Van Houdt et al., 2020) . . . 7
1.4 Architecture of a BiLSTM Model (Jatavallabha et al., 2024) . . . 8
1.5 Architecture of a BiGRU Model (Shiri et al., 2023) . . . 8
1.6 Architecture of a Attention Mechanism (Choi & Lee, 2023) . . . 9
1.7 Architecture of a Transformer Model (Choi & Lee, 2023) . . . 10
1.8 Architecture of a BERT Model (Rivera-Zavala & Martínez, 2021) . . . 11

2.1 Illustration of problem statement . . . 21
2.2 Proposed Framework Overview . . . 22
2.3 Presentation of emotion distribution in the GoEmotions dataset (Demszky et al., 2020) . . . 23
2.4 Traditional sentiment analysis . . . 25
2.5 Method of Feature Extraction . . . 25
2.6 Method of Fine-Tuning . . . 26
2.7 Fine Grained based sentiment analysis . . . 27
2.8 Hybrid Feature Fusion. . . . 28
2.9 Classification Setup . . . 29
2.10 confusion matrix . . . 30

3.1 Best-performing model per category (ML, DL, ATT) — F1-Weighted and MCC scores across the four datasets (FinBERT feature extraction) . . . 38
3.2 Approach 1 — Fine-tuning: best F1-Weighted and MCC scores per embedding (ML / DL / ATT) across four datasets . . . 40
3.3 Approach 2 — RoBERTa Feature Extraction (Dim=768): F1-Weighted and MCC scores of the best-performing model per category (ML, DL, ATT) across four datasets . . . 43
3.4 Best-performing model on 768+768 representation per category (ML, DL, ATT) — F1-Weighted and MCC scores across the four datasets (Concatenation Feature Extraction with Roberta-GoEmotion (27-768)) . . . 47
3.5 Best-performing model on 768+768 representation per category (ML, DL, ATT) — F1-Weighted and MCC scores across the four datasets (concatenation fine tuning FinBERT with Roberta-GoEmotion (27-768)) . . . 50
3.6 Best-performing model on 768+768 representation per category (ML, DL, ATT) — F1-Weighted and MCC scores across the four datasets (concatenation fine tuning RoBERTa with Roberta - GoEmotion (27-768)) . . . 52
3.7 Best-performing model on 768+768 representation per category (ML, DL, ATT) — F1-Weighted and MCC scores across the four datasets (concatenation fine tuning DeBERTa with Roberta-GoEmotion (27-768)) . . . 55
3.8 Best-performing model on 768+768 representation per category (ML, DL, ATT) — F1-Weighted and MCC scores across the four datasets (concatenation fine tuning XLM-RoBERTa with Roberta-GoEmotion (27-768)) . . . 57

3.9 Distribution of average emotion scores detected in Phrasebank 100% Dataset . 59
3.10 Distribution of average emotion scores detected in Phrasebank 50% Dataset . . 59
3.11 Distribution of average emotion scores detected in Gold Dataset 60
3.12 Distribution of average emotion scores detected in Twitter Dataset 60
3.13 Top 10 Emotions by Sentiment Class in the Phrasebank 100% Dataset 61
3.14 Top 10 Emotions by Sentiment Class in the Phrasebank 50% Dataset 61
3.15 Top 10 Emotions by Sentiment Class in the GOLD Dataset 62
3.16 Top 10 Emotions by Sentiment Class in the Twitter Dataset 62
3.17 Confusion Matrix of SVM (Best Model) in Approach 1 — Gold Dataset 65
3.18 Confusion Matrix of SVM (Best Model) in Approach 2 — Gold Dataset 66
3.19 Confusion Matrix of Self-Attention (Best Model) in Approach 3 — Gold Dataset 66
3.20 Confusion Matrix of XGBoost (Best Model) in Approach 1 — Phrasebank 50% Dataset 67
3.21 Confusion Matrix of -SVM (Best Model) in Approach 2 — Phrasebank 50% Dataset 67
3.22 Confusion Matrix of SVM (Best Model) in Approach 3 — Phrasebank 50% Dataset 68

List of Tables

1.1 Summary of Financial Sentiment Analysis Datasets 15
1.2 Summary of Related Works 17

2.1 Distribution of classes in the different datasets 25

3.1 Dataset Splitting Strategy for ML and DL Experiments 33
3.2 Hyperparameters Used for Classical ML Models 34
3.3 Hyperparameters Used for Deep Learning Models 35
3.4 Hyperparameters Configuration for Attention-Based Models 36
3.5 Hyperparameters Used for Fine-Tuning Models 36
3.6 Approach 1 — Features Extraction: FinBERT classification results across four datasets 37
3.7 Approach 1 — Fine tuning : 4 Models classification results across four datasets 38
3.8 Approach 2 — RoBERTa Embeddings: classification results across four datasets 41
3.9 Approach 3 — Concatenation Feature extraction: classification results across four datasets 44
3.10 Approach 3 — Concatenation Fine tuning FinBERT: classification results across four datasets 47
3.11 Approach 3 — Concatenation Fine tuning RoBERTa: classification results across four datasets 50
3.12 Approach 3 — Concatenation Fine tuning DeBERTa: classification results across four datasets 53
3.13 Approach 3 — Concatenation Fine tuning XLM-RoBERTa: classification results across four datasets 55
3.14 Top 5 Dominant Emotions per Sentiment Class across Datasets 63
3.15 Comparison with (Mir et al., 2025) across Datasets 68

List of Acronyms

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BiGRU	Bidirectional Gated Recurrent Unit
BiLSTM	Bidirectional Long Short-Term Memory
CatBoost	Categorical Boosting
CNN 1D	Convolutional Neural Networks 1 Dimension
DeBERTa	Decoding-enhanced BERT with Disentangled Attention
DL	Deep Learning
DT	Decision Tree
FinBERT	Financial BERT
FiQA	Financial Question Answering (Dataset)
FSA	Financial Sentiment Analysis
GPT	Generative Pre-trained Transformer
GRU	Gated Recurrent Unit
LightGBM	Light Gradient Boosting Machine
LLM	Large Language Model
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
RF	Random Forest
RNN	Recurrent Neural Networks
RoBERTa	Robustly Optimized BERT Approach
SVM	Support Vector Machine
XGBoost	EXtreme Gradient Boosting
XLM-R	XML-RoBERTa

General Introduction

In recent years, financial sentiment analysis has become essential for investors and institutions seeking actionable insights from news articles, earnings reports, and social media. Sentiment Analysis, a subfield of natural language processing (NLP) is used to automatically detect and classify emotional polarity (i.e., positive, negative, neutral) in a given text.

Early rule-based and lexicon-based approaches were simple to implement but ignored context, sarcasm, negation, and domain-specific language, leading to poor generalization. Machine Learning (SVM, Random Forest, Logistic Regression) improved performance by learning from labeled data. Deep Learning (LSTM, BiLSTM, GRU, CNN) further captured sequential dependencies and contextual information. Most recently, Transformers and attention mechanisms revolutionized the field by enabling models to focus on the most relevant parts of a text, with domain-adapted models like FinBERT bringing significant gains in financial NLP tasks .

Concurrently, traditional sentiment analysis frameworks exhibit critical structural limitations. By restricting analysis to coarse categorical polarities (positive / negative / neutral) traditional models fail to capture the underlying psychological states and behavioral dynamics that drive market movements. They inherently overlook the subtle psychological nuances and the diverse spectrum of human affect that govern financial decision-making.

To overcome these deficiencies, this research proposes the integration of Fine-Grained Emotion Modeling, leveraging comprehensive frameworks such as the GoEmotions taxonomy. Mapping distinct emotional dimensions (such as optimism, fear, surprise, or disappointment) enriches the semantic representation of financial texts, thereby providing the superior analytical granularity necessary for robust classification. Consequently, this study is driven and motivated by the following central research question:

Research Question: Do Fine-Grained Emotions Improve Financial Sentiment Classification?

The primary contribution of this work is the proposed framework. Diverging from isolated or strictly sequential methods found in the current literature, our methodological framework executes a feature fusion and concatenation process that couples representations from a highly specialized financial language model, such as FinBERT, with fine-grained emotion vectors extracted from a RoBERTa model pre-trained on the GoEmotions dataset. By leveraging the combined strengths of Machine learning models , deep learning architectures, and attention mechanisms, this work provides a concrete solution to bridge the gap between financial textual semantics and market psychology.

Overview of the Chapters

Chapter 1 — Background and Literature Review covers ML algorithms, DL architectures, attention mechanisms, transformer models, financial sentiment analysis, and fine-grained emotion modeling, concluding with a literature gap analysis.

Chapter 2 — Research Methodology and Proposed Approaches presents the datasets, emotion extraction pipeline, the three proposed approaches, and evaluation metrics .

Chapter 3 — Experiments and Results details the experimental setup, reports all empirical

findings, and provides cross-approach comparisons. Crucially, it evaluates the performance of the proposed frameworks against the core research questions. It concludes with an emotion-sentiment interaction analysis, a detailed error analysis, and a benchmarking comparison with existing literature.

Chapter 1: Background and Literature Review

Chapter 1

Background and Literature Review

1.1 Introduction

Artificial intelligence (AI) and natural language processing (NLP) have advanced in recent years, moving from traditional machine learning algorithms to deep learning models and transformer architectures. These advancements have enabled the development of powerful and efficient models, particularly within the financial domain. The first part of this chapter examines the theoretical underpinnings of AI, encompassing machine learning (SVM, Random Forest), deep learning, word embeddings, Transformer architectures, transfer learning, domain-adapted language models such as BERT, RoBERTa, and FinBERT. Based on these theoretical foundations, Part 2 applies these ideas to the practical field of Financial Sentiment Analysis (FSA). It examines the key characteristics of financial text (complex jargon, sarcasm, implied context, neutrality biais), the difference between binary sentiment and fine-grained emotion modeling using the GoEmotions taxonomy, and a review that enables the identification of current gaps. This shift between the two parties highlights how the theoretical tools presented in Part 1 will show as a foundation for a more nuanced hybrid approach that combines the financialization of FinBERT with the fine-grained recognition of emotions in order to overcome the limitations of traditional models.

Part 1: Background of AI

AI is one of the fields of study under computer science; its applications and uses can be found throughout different fields and disciplines. The focus of AI is to develop systems that can perform tasks that typically require human intelligence, such as problem solving, decision making, learning, perception, and natural language processing. The growth and progress of AI technology are dependent on advances in three areas: increased processing capabilities (hardware), developing better algorithms, and having access to very large data sets. Most important sub-domain of AI is:

1.2 Machine learning

Machine learning (ML) is a scientific discipline focused on developing mathematical algorithms and models capable of learning from data to perform specific tasks effectively. It aims to construct complex rules from data. ML can automate a wide range of tasks, particularly when people are unable to create manual instructions. These algorithms are employed across a variety of domains, including data mining, image processing, and predictive analytics. ML algorithms are broadly classified into four categories: supervised, unsupervised, semi-supervised, and

reinforcement learning (Saraswat & Raj, 2021). Examples of these algorithms include SVM, RF, LR, XGBoost, CatBoost, and LightGBM.

1.2.1 Support Vector Machine (SVM)

SVM is a model that uses a hypothesis space in the form of linear functions in a high-dimensional feature space, followed by an optimization-based learning method that integrates statistical learning bias. SVM classification aims to identify a boundary line that divides two distinct sets of data. The objective of SVM is to create an optimal boundary in order to create a maximum separation between the two data classes. SVM uses kernel algorithms to process high-dimensional datasets. SVM simply takes a few data points (Support Vectors) to create the model used in classification (Aditya et al., 2024).

1.2.2 Random Forest (RF)

RF which is based on the Decision Tree algorithm, is composed of numerous Decision Tree models trained on unique datasets with elements of randomization in the attribute-sets. These solutions offer advantages for performing predictions, particularly in terms of improved accuracy with incomplete records, resiliency against extreme values (outliers), optimized data storage capabilities, and feature selection methods that enhance Predictive Performance across larger and more complex datasets(Aditya et al., 2024).

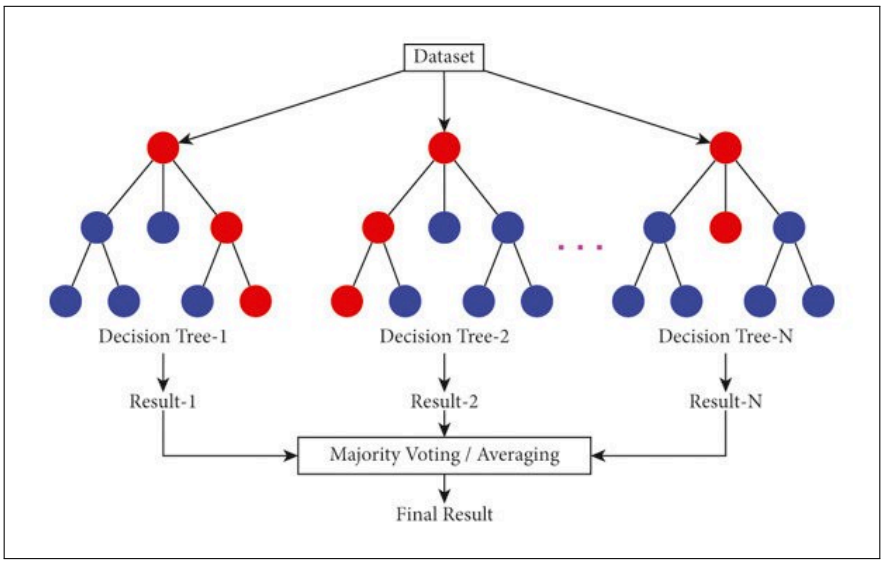


Figure 1.1. Architecture of a Random Forest Model (Khan et al., 2021)

1.2.3 Logistic Regression (LR)

LR is a type of statistical procedure used to analyze the relationship between 1 or more predictor (or independent) variables and a response or dependent variable that can take on either two or more values (Aditya et al., 2024).

1.2.4 eXtreme Gradient Boosting (XGBoost)

XGBoost is a gradient boosting algorithm, which builds strong predictive models by combining multiple weak learners (mainly decision trees). It improves performance by correcting errors iteratively and optimizing a loss function. Owing to its regularization techniques, it reduces overfitting and ensures generalization. XGBoost has a high rate of speed, can be processed

quickly, and is flexible, being able to process many different types of input data and provide a variety of output from traditional regression and classification models (Tomar et al., 2024).

1.2.5 Categorical Boosting Gradient (CatBoost)

CatBoost is a gradient boosting algorithm (short for Categorical Boosting). It leverages binary decision trees to form its base predictors and perform gradient boosting. Categorical features can be handled by CatBoost, and Bayesian estimators handle both ordered features and overfitting. The CatBoost method ranks the features in the created model using either Loss Function Change (LFC) or Prediction Values Change (PVC). The CatBoost machine learning model uses PVC by default. LFC is typically used to rank a particular model in relation to other models (Aditya et al., 2024).

1.2.6 Light Gradient Boosting (LightGBM)

LightGBM is a gradient boosting framework built on decision tree techniques is called LightGBM, or Light Gradient Boosting Machine. It is an effective application of Gradient Boosting Decision Trees (GBDT), in which performance is enhanced by sequentially training several trees. LightGBM is appropriate for large-scale data because of its speedy and memory-efficient design. Because of its good interpretability, accuracy, and efficiency, it is frequently employed in machine learning (Lokker et al., 2024).

1.3 Deep Learning

Deep Learning (DL) is an area specializing in using deep learning to automatically discover multiple levels of higher-level features in raw data by using layers of artificial neural networks. In contrast to traditional machine learning techniques, which rely primarily on human intervention for forming and extracting features manually, deep learning has the capability to perform feature learning and classification jointly from beginning to end. Such As, these models are better suited to operate on very large data sets and on complex patterns within unstructured data, such as speech or natural language (Shiri et al., 2023). Examples of these algorithms include CNN, RNN, LSTM, BiLSTM, GRU, and BiGRU.

1.3.1 Convolutional Neural Networks (CNN)

CNNs are deep learning models for tasks like computer vision and image classification that were first motivated by human visual perception. But CNNs-1D are moreover very successful in NLP. They eliminate the need for manual feature engineering by automatically learning features from data using essential components like convolutional layers, pooling layers, and fully connected layers. In NLP, word sequences are processed in place of pixels to extract local feature representations such as n-grams. CNNs combine feature extraction and classification into a two-stage architecture, as seen in figure 1.2, enabling the model to learn from beginning to end (Shiri et al., 2023).

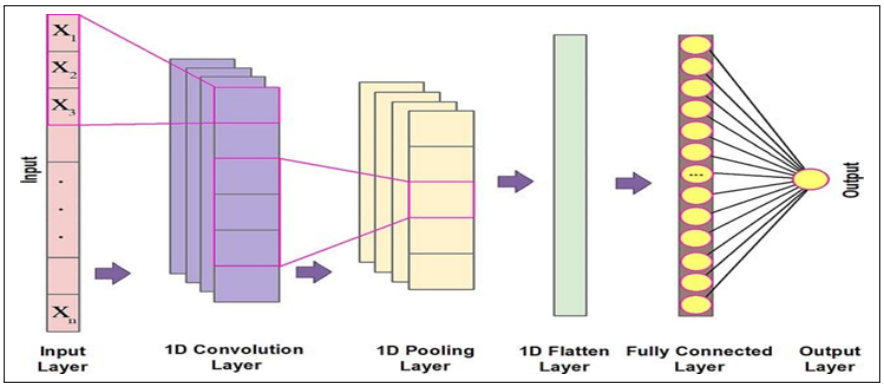


Figure 1.2. Architecture of a CNN-1D Model (Hoa et al., 2024)

1.3.2 Recurrent Neural Networks (RNN)

RNN is made to analyze sequential input by capturing temporal dependencies in internal memory. RNNs take input order into account, in contrast to conventional neural networks. However, short-term memory constraints plague basic RNNs, which prompted the creation of more sophisticated variations like LSTM and GRU (Shiri et al., 2023).

1.3.3 Long Short-Term Memory (LSTM)

LSTM is an advanced variant of recurrent neural networks, specifically designed to address the vanishing gradient problem inherent in standard RNNs and to capture long-term dependencies within sequential data. This is achieved through three gating mechanisms: the input gate, which controls the flow of new information into the cell state; the forget gate, which determines what information should be discarded from previous states; and the output gate, which regulates what portion of the cell state is propagated to the next time step. This architecture enables the selective retention and transmission of relevant information across arbitrarily long sequences, making LSTMs significantly more effective than standard RNNs for sequential modelling and text classification tasks. (Shiri et al., 2023).

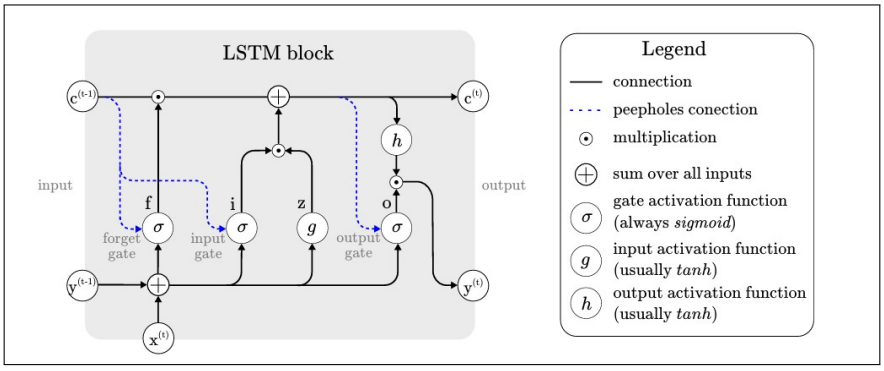


Figure 1.3. Architecture of a LSTM Model (Van Houdt et al., 2020)

1.3.4 Bidirectional LSTM (BiLSTM)

BiLSTM architecture extends the capabilities of LSTM by processing information in both forward and backward directions, thereby capturing contextual information from both past and future states. This characteristic improves performance across various sequential data tasks(Shiri et al., 2023).

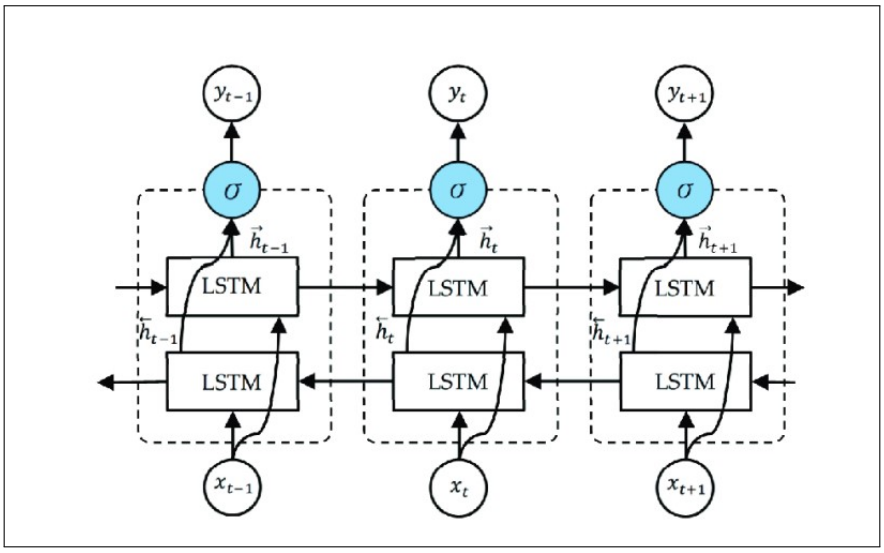


Figure 1.4. Architecture of a BiLSTM Model (Jatavallabha et al., 2024)

1.3.5 Gated Recurrent Unit (GRU)

GRU is a streamlined version of the Long Short-Term Memory (LSTM) that captures crucial dependencies in sequential data while increasing speed and efficiency by incorporating gates into update and reset procedures.(Shiri et al., 2023)

1.3.6 Bidirectional GRU (BiGRU)

By combining both forward and backward processing, bidirectional GRU improves this model even more and enables a more thorough comprehension of sequence context (Shiri et al., 2023).

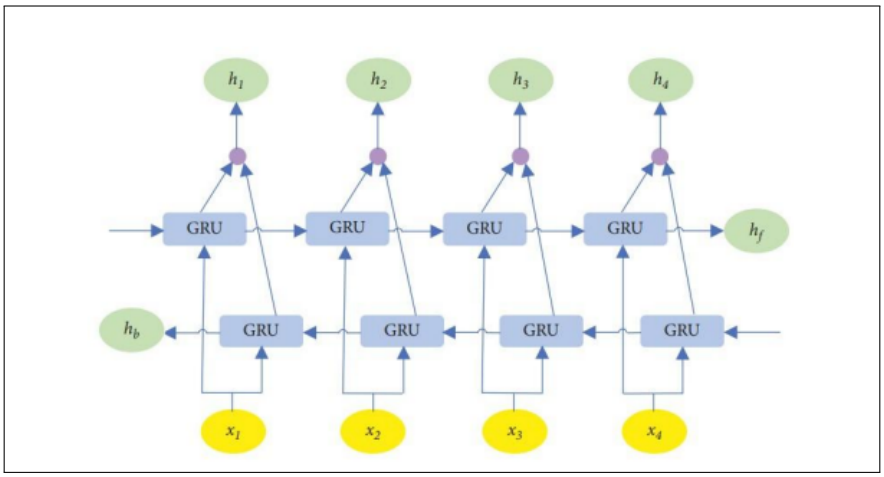


Figure 1.5. Architecture of a BiGRU Model (Shiri et al., 2023)

1.4 Word Embeddings to Contextual Models

Embedding techniques in NLP have evolved from traditional static embeddings to advanced contextualized models. Static embeddings are fixed mathematical vector representations of words that do not change according to context. Early approaches used 'word embeddings' as basic representations of words, but they lacked context. The use of 'contextual embeddings', as are used in BERT and GPT, has changed that, allowing for dynamic representations that can change based on the words in the context where they occur. As a result of this change, performance in tasks like sentiment analysis, translation, and question/answering has improved (Alkaabi et al., 2025).

1.5 Attention Mechanisms

Attention is introduced in sequence-to-sequence models to address the limitations of recurrent neural networks. It enables the evaluation of the relative contribution of each element within a sequence. This process relies on Query (Q), Key (K), and Value (V) representations, where attention scores are computed between Q and K, normalized through a softmax function, and applied to weight the values (V), producing a context-dependent representation (Choi & Lee, 2023).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{1.1}$$

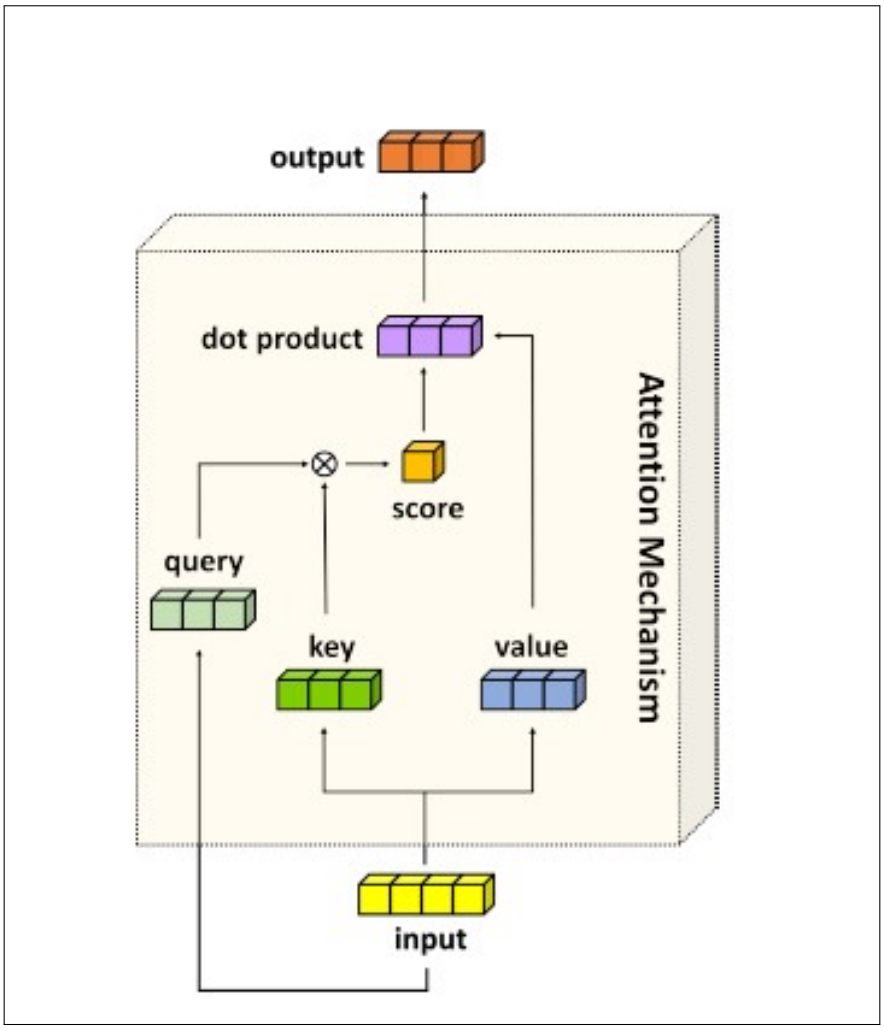


Figure 1.6. Architecture of a Attention Mechanism (Choi & Lee, 2023)

1.5.1 Self-attention

Self-attention (or Intra-attention) uses an attention mechanism that compares and relates the items from one sequence in order to create an overall representation of that sequence. All of the Queries (Q), Keys (K), and Values (V) used in this process will come from the same sequence, so that each item will compare itself to and pay attention to all other items of that same sequence (Vaswani et al., 2017).

1.5.2 Multi-Head Attention

Multi-Head Attention extends this concept by projecting the Queries (Q), Keys (K), and Values (V) multiple times into different, lower-dimensional representation subspaces which constitute the ”heads” . This process allows the model to compute multiple (or parallel) attentional mechanisms instead of just one over all sequences. Each independent representation from the heads is then combined together via an additional linear projection into richer, multi-dimensional representations (Vaswani et al., 2017).

1.5.3 Hierarchical Attention

A hierarchical attention mechanism designed to reflect the hierarchical nature of documents. It uses two separate attention mechanisms, one at both the word-level and the sentence-level, which allow the model to apply attention differentially depending on the relative importance of the items in the document when constructing its representation of the document(Yang et al., 2016).

1.6 Transformer Architecture

The Transformer is a sequence modeling architecture based on self-attention mechanisms. It follows an encoder–decoder framework in which both components are composed of stacked identical layers. Each encoder layer includes a multi-head self-attention mechanism and a position-wise feed-forward network, combined with residual connections and layer normalization. The decoder extends this structure by incorporating a masked self-attention mechanism and an encoder–decoder attention layer. The model processes input sequences in parallel and relies on positional encoding to represent the order of elements. Its design enables the modeling of dependencies between sequence elements without relying on recurrence (Choi & Lee, 2023).

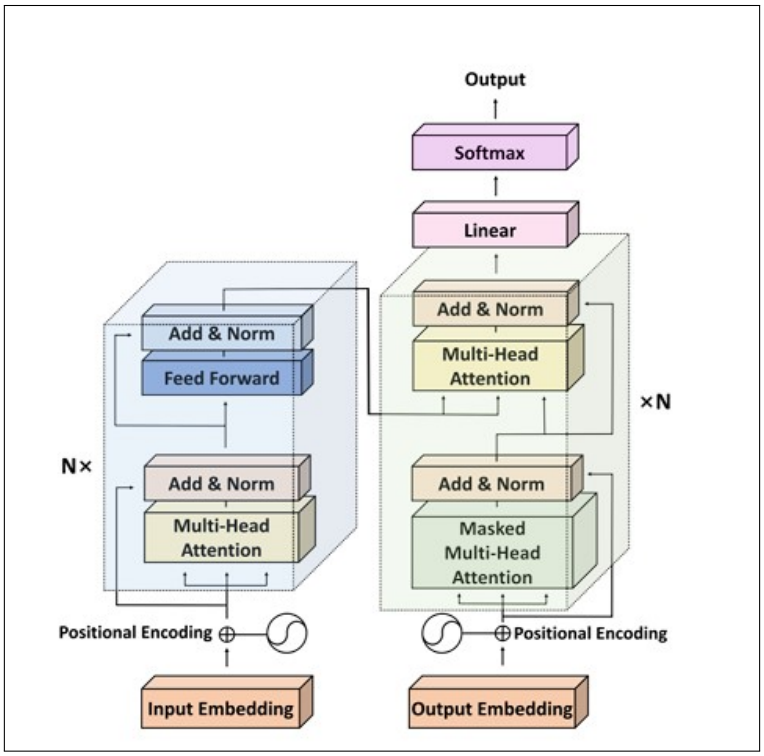


Figure 1.7. Architecture of a Transformer Model (Choi & Lee, 2023)

1.6.1 Transfer Learning in NLP

In natural language processing, transfer learning entails pre-training a model on a sizable unlabeled corpus for a generic language modeling goal and then fine-tuning the resulting parameters on smaller labeled datasets to adapt them to one or more downstream tasks. This method transfers pre-training information to solve data scarcity in domain-specific applications.

1.6.2 Pre-trained Language Models

A. **BERT** constitutes a language representation model that pre-trains deep bidirectional representations from unlabeled text through Transformers (Devlin et al., 2019).The model receives pre-training on a data corpus that includes the BookCorpus, which contains 11,038 unpublished books from 16 genres, and text passages from English Wikipedia. BERT employs bidirectional contextual modeling, whereby word representations vary according to sentence context. The pre-training phase replaces a portion of tokens with the [MASK] symbol and reconstructs the original tokens through the auto-encoding language modeling objective. The training incorporates two tasks: Masked Language Modeling and Next Sentence Prediction (Rajapaksha et al., 2021).

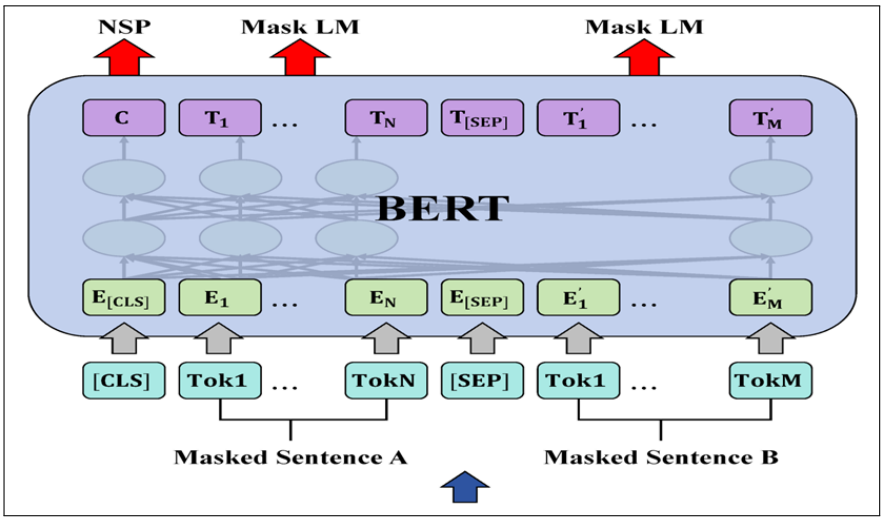


Figure 1.8. Architecture of a BERT Model (Rivera-Zavala & Martínez, 2021)

B. **RoBERTa** maintains the same architecture as BERT. The model receives pre-training on a larger data corpus than BERT. The modifications in RoBERTa encompass dynamic masking patterns applied to each training instance, training on longer sequences, and the removal of the next sentence prediction task (Liu et al., 2019). The training dataset undergoes duplication 10 times, which enables each sequence to receive masking in 10 different ways across 40 epochs. RoBERTa achieves performance improvements that range from 2% to 20% over BERT on benchmark tasks (Rajapaksha et al., 2021).

C. **XLM-RoBERTa** is a large-scale multilingual language model based on the Transformer architecture. It is trained using the RoBERTa approach (Masked Language Modeling) on a massive dataset covering 100 different languages, using more than 2 terabytes of filtered Common Crawl data (Conneau et al., 2020).

D. **DeBERTa** uses the same underlying architecture Like BERT and RoBERTa. however, it introduces two new features to increase performance. First, it uses Disentangled Attention. In this mechanism, each token is represented by two separate vectors: one for content and one for relative position. Second, it employs an Enhanced Mask

Decoder. This decoder incorporates absolute position information during the pre-training phase. These improvements allow Debert to better model both syntactic and semantic relationships between words compared to RoBERTa, providing superior results with less training data (He et al., 2020).

1.6.3 Domain-Specific Language Models

This work is based on two specific models: FinBERT and RoBERTa.

A. **FinBERT** is a domain-specific adaptation of the BERT architecture designed to overcome the limitations of general-purpose language models in processing financial text. It utilizes a two-step transfer learning approach, starting with unsupervised pre-training on a massive financial corpus to master domain-specific vocabulary and semantics, followed by target-task fine-tuning where a classification head is added to the [CLS] token.(Araci, 2019).

B. **SamLoweroberta-base-go-emotions** is a RoBERTa variation fine-tuned for multi-label emotion classification, made publicly available on the Hugging Face model hub (Lowe, 2026). The model uses RoBERTa-base weights and is further trained on the GoEmotions dataset, a large-scale corpus of 58,000 Reddit comments annotated with 28 emotion categories, in order to acquire the affective and contextual relationships characteristic of fine-grained emotion recognition. The model produces 28 probability scores as output, one for each emotion label, and a threshold of 0.5 is typically applied to binarize the predictions. This multi-label design allows the model to assign one or multiple emotion labels simultaneously to a given input text, making it well suited for tasks requiring nuanced emotional understanding beyond simple polarity classification .

Part 2: Financial Sentiment Analysis and Emotion Modeling Foundations

1.7 Financial Sentiment Analysis (FSA)

1.7.1 Definition

Financial Sentiment Analysis (FSA) is defined as a field of study that analyzes people's sentiment from financial texts, measures and quantifies investor sentiment, and is ultimately grounded in the applications of market prediction and financial decision-making. It is a domain-dependent "suitcase problem" that specifically studies investor sentiment and financial textual sentiment (Du et al., 2024).

1.7.2 Specificities of the Financial Domain

The defining characteristics that differentiate Financial Sentiment Analysis (FSA) from generic sentiment analysis are; the terminology used in FSA is highly specialized (using a different set of terms with specific meanings for each term), making it necessary to have familiarity with these terms prior to working with them. It is common to see the use of figurative language and rhetorical devices within the financial industry to describe market conditions. Because of the emphasis on the accuracy of conveying information with little detail, specialists in the FSA field also tend to employ very concise terminology in order to effectively convey very detailed

information in as few words as possible, therefore requiring models that can decode sentiment from highly condensed sentence structures. Additionally, since the upward or downward movement of a particular sentiment influences the sentiment meaning itself, directionality is an feature of FSA; Lastly, another defining characteristic of FSA is that it often includes quantitative data as well as qualitative text in one form of financial communication, therefore requiring a model to assess both quantitative and qualitative variables to obtain an accurate representation of the entire financial market in context (Du et al., 2024).

1.7.3 Objectives of FSA

The targets of Financial Sentiment Analysis (FSA) are cut up into number one studies streams, with an overarching purpose of helping monetary markets.

A. **Technique-Driven Objectives** The primary objective of this research stream is to enhance the performance of FSA systems through the design of novel techniques, models, and domain-specific algorithms. High-quality human-annotated datasets are developed to support the training and evaluation of these methods. A further objective involves refining the granularity of sentiment analysis by improving the accuracy with which models detect the degree of sentiment polarity within financial documents. To this end, analysis can be conducted at multiple levels of textual granularity, including paragraph-level, sentence-level, and targeted aspect-based sentiment analysis (Du et al., 2024).

B. **Application-Driven Objectives** This stream leverages financial sentiment analysis for downstream applications within financial markets. Key areas of focus include analyzing sentiment signals to capture market dynamics, examining the correlational relationships between investor sentiment and market behavioral shifts, and integrating sentiment measures into financial decision-making and business strategy formulation. A further concern is the identification of appropriate real-world deployment contexts for existing sentiment analysis methodologies (Du et al., 2024).

1.7.4 Key Challenges

Financial Sentiment Analysis presents several inherent challenges that distinguish it from general-purpose sentiment classification. Domain-dependent neutrality bias renders terms such as ”liability” or ”debt” semantically neutral in financial contexts, undermining the applicability of general-purpose lexicons. The frequent use of sarcasm, irony, and rhetorical devices further complicates automatic sentiment detection, as does the reliance on implicit contextual knowledge not explicitly stated within the text. Finally, the construction of high-quality annotated corpora remains a persistent bottleneck, requiring substantial domain expertise and yielding variable levels of inter-annotator agreement (Du et al., 2024).

1.8 Landscape of Financial Sentiment Datasets

The degree to which NLP can be utilised effectively in finance depends largely upon the quality of the training data used. This section outlines previously used benchmark resource categories, and presents domain-specific new proprietary datasets specifically targeted for use in commodity markets.

1.8.1 Benchmark Research Datasets

A. **Financial PhraseBank** constitutes as a gold-standard benchmark in financial NLP, contains over 4,840 phrases collected from financial news publications and from press

releases issued by various companies. This dataset was created specifically for training and evaluating sentiment models and is used to classify the sentiment of a piece of text into three classes (positive, negative, neutral) based upon how it will potentially impact stock prices. The one main strength of the Financial PhraseBank is that it has been developed using a high degree of reliability through the use of a meticulous annotation process involving 16 finance and economics experts. In addition to this, the Financial PhraseBank varies in terms of the level of consensus that was achieved by the 16 annotators for each sample, providing four different types of consensus levels (50%, 66%, 75%, 100%) (Malo et al., 2014).

B. **FiQA Sentiment Dataset** Data from microblogging and financial news is included in the FiQA dataset. Each input is given a continuous sentiment score from -1 to +1. This allows models to do regression tasks to understand the exact amount of sentiment rather than being limited to simple categorical classifications (Maia et al., 2018).

C. **SEntFiN (Entity-Aware)**Approximately 10,700 financial news headlines make up the SEntFiN (Entity-Aware) dataset. It differs from other datasets in that each company is given a sentiment (Positive, Negative, or Neutral) within a single headline, providing the ability to accurately model complex news headlines where different companies could be affected by differences in the economy at the same time.

1.8.2 Social Media Datasets

A. **Twitter Financial News** is a carefully compiled set of Twitter microblogging financial data, openly accessible on Kaggle (Costa, 2026). These short, quick, impactful tweets are typically placed into three types of sentiment (Positive, Negative, Neutral) in order to understand how breaking news and global discussions around finance can influence the movement of the markets.

B. **StockTwits** includes a ton of data from a social network for investors. It can be looked at in different ways featuring; cashtags which are a type of stock ticker (like $GOLD), user-created labels and a way for people to express their emotion about stocks by labeling them as (bullish or bearish). This dataset is mainly used to show what real-time retailers feel about stock prices and the discourse affecting marketplace volatility (Li et al., 2025).

C. **FinSen** dataset is a large-scale resource comprising 160,000 economic and financial news records from 197 countries, spanning the years 2007 to 2023. Each entry features automated sentiment scores generated by FinBERT and is paired with S&P 500 stock market data (Liang et al., 2024).

D. **Reddit WallStreetBets (WSB)** is a database of all posts and comments taken from this subreddit. The aim of the data set is to categorize market sentiment for speculative trading into Bullish or Bearish classes. The WSB data set will provide insight into how the viral nature of the social media conversation impacts the trading of meme stocks (Gianstefani et al., 2022).

1.8.3 Sentiment Analysis in the Gold Commodity Market

Gold Dataset has multiple columns, including a column for labeled news headlines with the label ”Price Sentiment” and other features such as the date of publication, source URL of the news story, and the raw content of the news. There are a total of 10,600 gold related news headlines in the dataset, for the dates from 2000 to 2019, indicating whether the news headline expressed a price increase (positive), price decrease (negative), price stability (neutral), or no direction of price at all (none) (Sinha & Khandait, 2021).

1.8.4 Synthesis of Data Resources

To conclude this section, the table 1.1 summarizes the datasets discussed:

Table 1.1. Summary of Financial Sentiment Analysis Datasets

Dataset	Type	Number of Lines	Number of Columns
Financial PhraseBank	News	4,840 sentences	2 columns (Sentiment, Sentence)
FiQA	News/Blogs	1,174 sentences	4 columns (Sentence/Snippet, Target, Aspect, Sentiment Score)
SEntFiN	News Headlines	10,753 headlines	Multiple columns (Headline, Entity, Sentiment Label)
Twitter Financial News	Social Media	11,932 tweets	2 columns (Text, Label)
StockTwits	Social Media	∼1.5 million records	Multiple columns (Text, Symbol, Sentiment, Timestamp)
FinSen	News/Economic Data	160,000 records	Multiple columns (Text, Sentiment, S&P 500 data)
Reddit (WSB)	Forum	∼1.5 million records	Multiple columns (Post/Comment, Score, Timestamp, Sentiment)
Gold Dataset	Commodity	10,570 headlines	Multiple columns (Price Sentiment, Dates, URL, News content, etc.)

1.9 Fine-Grained Emotion Modeling

1.9.1 Comparison between Sentiment Analysis and Emotion Polarity

Affect is reduced by conventional sentiment analysis to a polarity classification with three categories: neutral, negative, and positive. The method uses lexicons to detect contextual polarity inside a sentence or document after assigning prior polarity to words or phrases. This method produces a one-dimensional, scalar representation of affect.

A finer granularity is used in discrete emotion modelling. It divides affect into discrete categories, including fear, wrath, sadness, joy, and disgust. The approach takes into account the particular emotion type, it is intensity, contextual triggers, related behaviours, and situational antecedents. It is based on psychology theories of basic emotions.

The key distinction is in granularity: discrete emotions create a categorical representation that is richer and more accurate to subjective experience, while polarity reduces affect to a single dimension. A major obstacle for sophisticated sentiment analysis systems and emotion ontologies is this move toward fine-grained modelling (Park & Storey, 2023).

1.9.2 GoEmotions Taxonomy

The GoEmotions dataset represents a large-scale corpus designed for fine-grained emotion classification, comprising 58,000 English Reddit comments. This framework extends beyond basic sentiment polarity by incorporating 28 emotion categories, including 27 distinct emotions without a neutral label. These categories encompass a wide spectrum of affective states, including positive emotions such as admiration and gratitude, negative states like remorse and disappointment, and ambiguous categories such as realization or surprise. The hierarchical

organization of the taxonomy allows for the differentiation of emotions based on intensity and sentiment similarity (Demszky et al., 2020).

1.10 Related Work

This section reviews the literature underpinning this work, organized into two areas: financial sentiment analysis and fine-grained emotion recognition. The first traces the evolution from lexicon-based methods to modern transformer and LLM-based models. The second examines advances in detecting nuanced and underrepresented emotions through knowledge-enriched and personalized architectures.

1.10.1 Financial Sentiment Analysis

The development of financial sentiment analysis (FSA) has gone through four principal techniques.

Lexicon-based were among the first methods, depending on predetermined financial dictionaries like the one put forth by (Loughran & McDonald, 2011). Despite being widely used, these approaches frequently misclassify financial terms because they rely solely on word counting and do not capture contextual meaning.

Machine learning models such as Support Vector Machines (SVM) and Naive Bayes (NB), combined with TF-IDF feature extraction, offered notable improvements; (Roldan, 2023) reported accuracies of 74% for SVM and 71% for NB on the Financial PhraseBank dataset, though these approaches still require manual feature engineering and struggle with semantic dependencies.

Deep learning models including CNN, LSTM, and BiLSTM further advanced the field by automatically learning text representations: (Cheng & Chen, 2021) combined FinBERT with BiLSTM and a multi-head attention mechanism, realizing 79% accuracy on Chinese financial texts, while (Cicekyurt & Bakal, 2025)showed that augmenting training data using FinBERT-based pseudo-labeling improved F1-scores by up to 20%.

Transformer and LLM-based models nowadays, the The state of the art at the moment is represented by transformer and LLM-based models. The FinBERT model, created by (Araci, 2019), adapted BERT to the financial domain through further pre-training on the TRC2-financial corpus, achieving 97% accuracy on Financial PhraseBank sentences with full annotator agreement and 86% on the complete dataset, and greatly outperformed all earlier techniques on every measured metric. (Zhang et al., 2024) highlighted the value of fine-grained (aspect-level) sentiment analysis for enhanced financial distress prediction , being the first to integrate such features into financial distress prediction (FDP) models and demonstrating improved early warning capabilities. (Mir et al., 2025) introduced a joint topic-emotion modeling framework leveraging the Neural Variational Document Model (NVDM) for topic discovery (coherence: 0.68, perplexity: $-$ 8.5) combined with a fine-tuned FinBERT for emotion classification, achieving 96.3% accuracy, 97.8% F1-score. This integrated approach links latent financial topics with nuanced investor emotions, enhancing interpretability for market analysis and trading strategies. (Sun et al., 2025) proposed EnhancedFinSentiBERT, incorporating dictionary knowledge and a neutral feature extractor, reaching up to 97.6% F1-score; and most recently, (Shi et al., 2025) introduced FSAExpert, which fuses internal keyword knowledge with LLM-generated external knowledge through a chain-like reasoning mechanism, achieving up to 99.6% accuracy.

1.10.2 Fine Grained Emotion Recognition

Fine-grained emotion recognition has progressed due to extensive datasets and contextual modelling.(Zhong et al., 2019)combined emotion lexicons with ConceptNet(is a large-scale

semantic network) knowledge graphs in Knowledge-Enriched Transformers (KET) to detect emotions in conversations. (Suresh & Ong, 2021) proposed Knowledge-Embedded Attention (KEA) with ELECTRA(is a pre-trained language model that uses replaced token detection instead of masked language modeling), and NRC-VAD(is an emotion lexicon that associates words with three emotional dimensions: Valence, Arousal, and Dominance) lexicons. It achieved 54.1% top-1 accuracy on EmpatheticDialogues (32 emotions) (Rashkin et al., 2019) and reduced confusion between closely related emotions such as "afraid" and "terrified". (Kocoń et al., 2021) demonstrated that personalized architectures outperform majority voting, which may lead to culturally biased models. (K. Wang et al., 2024) showed that using ProtAugment to augment minority categories (grief, pride, nervousness, and relief) improved the macro-F1 score on GoEmotions (28 categories, 58k Reddit comments) from 0.46 to 0.52, while also highlighting limitations of LLMs such as hallucination and over-labeling.
In 2025,(Ngo & Kocoń, 2025) demonstrated that Sequence Sentence Classification (SSC) using DistilRoBERTa surpasses sentence-level baselines. Their personalized UserID approach improved the micro F1-score by 15.4 percentage points, particularly for underrepresented emotions such as fear, surprise, and anticipation. (Alqarni et al., 2025) proposed Emotion-Aware RoBERTa by incorporating an Emotion-Specific Attention (ESA) layer and TF-IDF gating, improving the accuracy by 10.87 percentage points over the baseline RoBERTa and achieving 96.77% accuracy. (Ren et al., 2025) introduced Emotion In-Context Learning (EICL), which uses emotionally analogous examples and a two-stage exclusion strategy, outperforming conventional ICL across five LLMs and four datasets. (Tong et al., 2025) proposed EmoSync, A multi-stage framework combining unimodal reasoning, chain-of-thought self-correction, and ensemble-of-perspectives to enhance emotion recognition accuracy.

Table 1.2. Summary of Related Works

N	Work	Datasets	Approche	Result
1	(Zhang et al., 2024)	**CNRDS** (Chinese Research Data Services) **&** **CSMAR** (China Stock Market & Accounting Research)	Unsupervised fine-grained sentiment extraction using Attention-Based Aspect Extraction combined with emotion dictionaries, and Validation of the extracted features on ML and DL models including SVM, DT, LR, CNN, and DNN.	Best result with Decision Tree: Accuracy = 92.77%, F1-Macro = 92.68%

Continued on next page...

Table 1.2 — *(continued)*

N	Work	Datasets	Approche	Result
2	(Mir et al., 2025)	**FinSen & StockEmotions**	Neural Variational Document Model NVDM (topic modeling) + Fine-tuned FINBERT (emotion classification) — joint multi-task learning with feature fusion.	Accuracy = 96.3%, F1-score = 97.8%
3	(K. Wang et al., 2024)	**GoEmotions & CARER**	fine-tuned BERT/RoBERTa on GoEmotions, enhanced with ProtAugment on minority classes and CARER transfer learning, outperforms zero-shot GPT-4.	It improved accuracy by up to 34% and F1-score by up to 12%.

1.10.3 Literature Gap

Key benchmarks for sentiment analysis (e.g., Twitter Financial News, Financial PhraseBank, Gold) have focused primarily on assigning one of three labels (e.g., positive, negative, neutral), and thus do not capture the diversity of emotional states that are implied by these labels. On the other hand, fine-grained emotion taxonomies such as GoEmotions provide 27 discrete emotional states but do not explicitly associate these states with a polarity (e.g., positive/negative), thereby rendering it impossible to use a single label system. To bridge this gap, we propose a hybrid framework that integrates polarity detection with emotional classification. By sequentially determining the sentiment orientation and then isolating the specific emotions consistent with that polarity, this dual-output architecture provides both a macroscopic affective view and microscopic emotional insights for downstream applications.

1.11 Conclusion

In this chapter, we presented the foundational principles of AI and NLP, establishing their direct connection to Financial Sentiment Analysis . In the first part, we traced the evolution of AI from traditional statistical Machine Learning algorithms to advanced Deep Learning models, leading up to attention mechanisms and modern Transformer architectures. Furthermore, we examined the impact of transfer learning and pre-trained language models, with a specific focus on domain-specific adaptation for financial applications. In the second part, we explored the foundational landscape of financial sentiment, detailing its unique domain challenges and benchmark datasets—including social media and commodity markets. Finally, we addressed the paradigm of fine-grained emotion modeling through the GoEmotions taxonomy, concluding with a comprehensive review of related work to identify current gaps in the literature.

Chapter 2: Research Methodology and Proposed Approaches

Chapter 2

Research Methodology and Proposed Approaches

2.1 Introduction

This chapter outlines the research methodology that is utilized in order to enhance financial sentiment analysis through the acquisition of emotional signals with a fine grained. Most current or traditional models fail to identify emotional signals that influence the market at a subtle level, this study proposes a framework that combines domain-specific semantics with high-resolution emotional features. The proposed methodology consists of preprocessing multiple datasets and employing three distinct modeling strategies: Domain-Specific Modeling, Emotion-Driven Modeling, and Hybrid Feature Fusion. The framework aims to increase classifier performance and provide more accurate and provide more accurate and actionable sentiment analysis through the vector concatenation of financial and emotion vectors on a common scale.

2.2 Problem statement

The majority of current models in the field of financial sentiment analysis rely on coarse-grained classification, which generally divides data into three major classes: neutral, negative, and positive. However, financial discourse is characterized by extreme nuance, where the intensity of a statement often carries more weight than its mere polarity. For instance, a ”slight decline” and a ”market crash” are both negative, but they trigger fundamentally different financial decisions. The central problem of this research is that simplistic classification may oversimplify the semantic richness of financial texts, leading to a loss of critical information. This study aims to bridge this gap by investigating the impact of fine-grained analysis.

Research Question Do Fine-Grained Emotions Improve Financial Sentiment Classification?

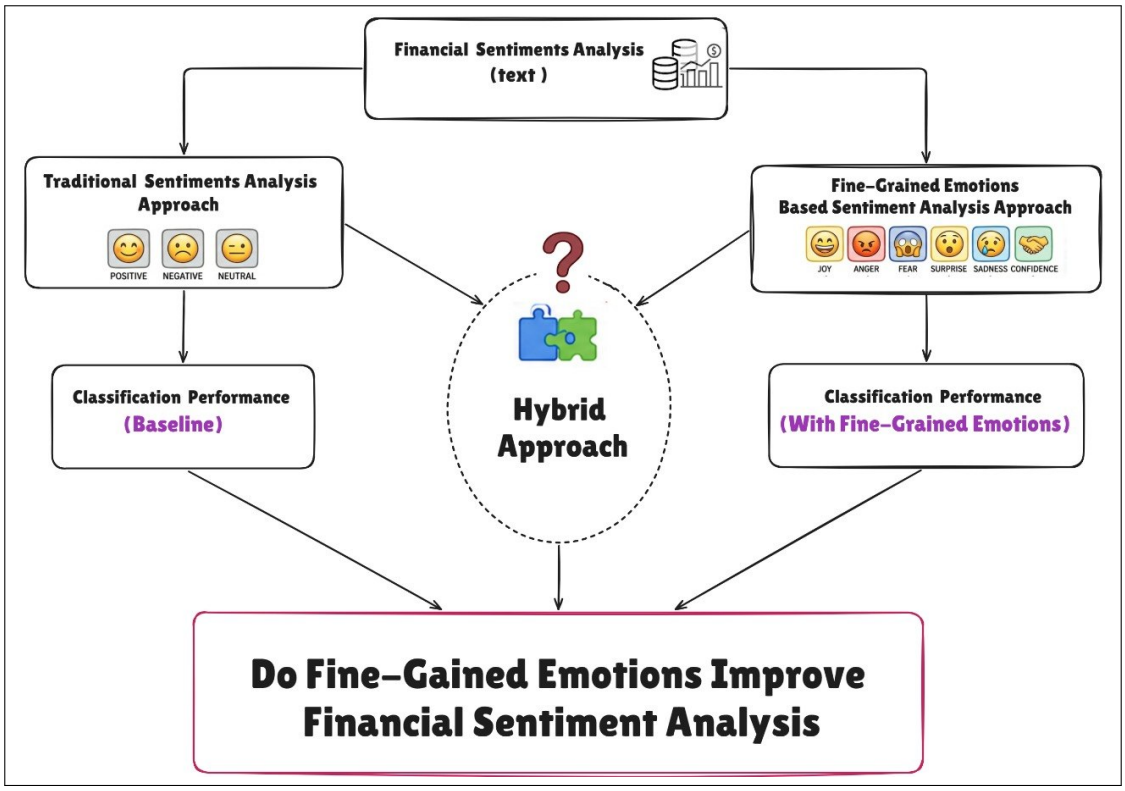


Figure 2.1. Illustration of problem statement

2.3 Proposed Framework Overview

The proposed framework analyzes financial sentiment through three main approaches. The first approach exploits the domain context through two strategies: feature extraction using FinBERT to generate fixed financial embeddings, and fine-tuning of pre-trained language models including FinBERT, RoBERTa, DeBERTa and XLM-RoBERTa, to adapt them specifically to financial sentiment classification. The second approach captures the emotional context by using RoBERTa pre-trained on the GoEmotions dataset to extract fine-grained emotional signals from financial texts. The third approach performs a fusion of both financial and emotional representations via Hybrid Feature Fusion. All three approaches share the same classification pipeline comprising ML, DL, and Attention-based models, and are evaluated to validate whether combining financial and emotional representations improves sentiment classification performance.

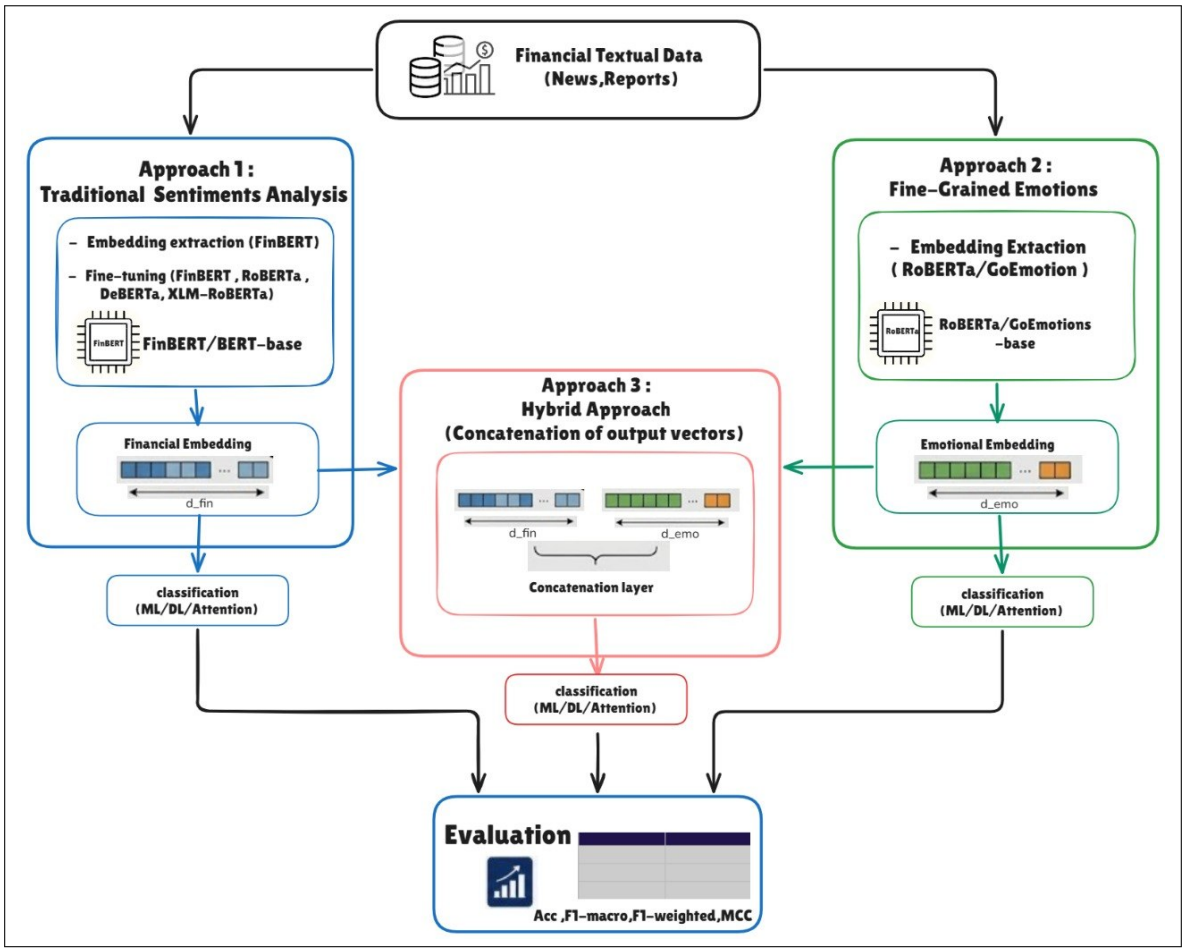


Figure 2.2. Proposed Framework Overview

2.4 GoEmotions dataset

In the GoEmotions dataset, the neutral class is excluded because it is highly dominant in the overall data distribution. Therefore, this study focuses only on the 27 emotion categories to ensure a more balanced and fine-grained emotional analysis. The dataset contains approximately 58,000 Reddit comments annotated with 27 emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, and surprise.

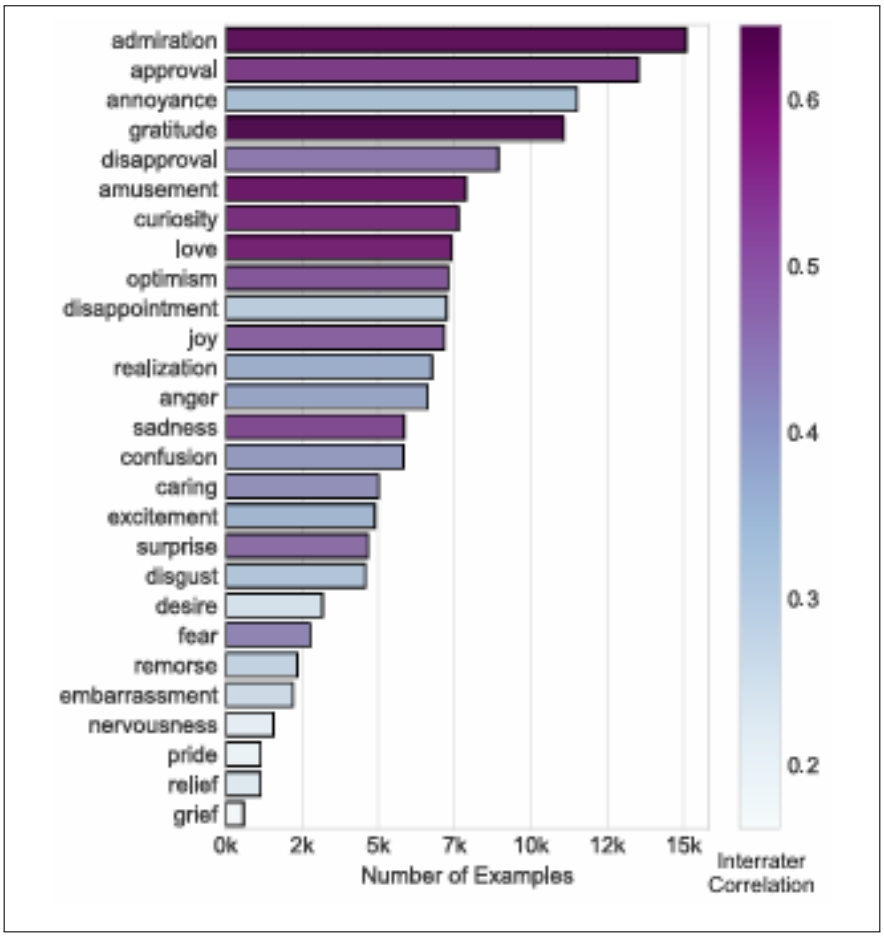


Figure 2.3. Presentation of emotion distribution in the GoEmotions dataset (Demszky et al., 2020)

The dataset is important because it enables modelling of subtle and complex emotional representations, not just simple sentiment polarity. Its fine-grained taxonomy gives language models a greater ability to distinguish subtle emotional signals within text. This is especially relevant in financial sentiment analysis, where market behaviour can be better explained by emotions such as fear, optimism, uncertainty or disappointment than by coarse-grained polarity labels. By integrating GoEmotions features in financial datasets, we can assess if more detailed emotional representations improve trend identification, sentiment classification, and predictive performance in financial markets (Demszky et al., 2020).

2.5 Financial Sentiment Datasets

One crucial stage in assessing how fine-grained emotions affect the classification of financial sentiment is the data selection process. The technical features and class distributions of the four different data sources that are used in the project design are described below.

2.5.1 Financial PhraseBank (50%, 100% agreement)

The Financial PhraseBank dataset available at Kaggle platform available at Kaggle platform. It was created by Pekka Malo, Ankur Sinha, Pyry Takala and colleagues from the Aalto University School of Business. It is a widely used benchmark dataset for financial sentiment analysis, which contains financial sentences extracted from press releases, company reports and financial news, and these sentences are manually annotated into positive, negative and neutral classes by finance experts (Malo et al., 2014).

The dataset is distributed in four agreement levels 50%,66%,75% and 100% agreement . In this study, the 50% and 100% agreement settings are utilized. The 50% agreement version has 4840 samples (2128 neutral, 1852 positive, and 860 negative) and has more ambiguous and challenging annotations. The 100% agreement version comprises 2264 samples (827 neutral, 987 positive and 450 negative) and represents a reliable gold-standard subset with full annotator consensus (Malo et al., 2014).

The Financial PhraseBank is a popular choice for evaluation of transformer-based financial language models such as FinBERT and RoBERTa due to its domain-specific vocabulary and high-quality annotation process.This dataset has been used in several studies for financial sentiment analysis (Araci, 2019), stock prediction (Kalyani et al., 2016) and transformer robustness evaluation (R. Wang et al., 2025).

2.5.2 Financial Sentiment Analysis GOLD Headlines

The GOLD Market Headlines dataset, available on Hugging Face platform, was curated and published by Saguaro Capital Management, LLC. It is a financial sentiment corpus composed of stock market headlines and technical commentary used for gold market analysis tasks. The dataset contains 10570 rows before preprocessing, with an initial distribution of approximately 4412 positive samples, 3814 negative samples, 376 neutral samples, and 1968 "None" samples. The "None" class includes headlines that are not directly related to financial sentiment or price movement signals (Sinha & Khandait, 2021).

To address class imbalance and improve data consistency, the GOLD Market dataset is reformulated into a binary classification task by removing the neutral (376) and "None" (1968) classes, resulting in a final dataset of 8226 samples . This preprocessing step helps stabilize the data distribution and focuses learning on clear positive and negative market signals.

This dataset is significant because it captures the technical and colloquial language used by stock market analysts . Filtering the neutral and unrelated classes reduces noise during the learning of emotional vectors by concentrating on explicit directional market signals, which improves algorithm convergence. This method helps evaluate whether the emotions recognized by GoEmotions improve gold market identification (Sinha & Khandait, 2021).

2.5.3 Twitter Financial News

The Twitter Financial News dataset, which is available on the Kaggle platform, was created by the dataset contributors and constructed using data collected through the Twitter (X) API, where financial messages were extracted and annotated for sentiment analysis tasks. The dataset contains11927rows spread over two columns: Text and Label. The distribution of the class includes approximately 7740 positive samples, 1789 negative samples, and 2398 neutral samples (Adams et al., 2023). This structure is consistent with the terse, information-rich style of financial messages that originate on social media platforms.

This dataset is essential to assess the performance of the proposed hybrid approach on data with high emotional and contextual load . This shows that financial sentiment is not just about polarity classification but about richer emotional dimensions such as optimism, fear and uncertainty, which can be captured by the GoEmotions model. The analysis of this dataset demonstrates the flexibility of the proposed methodology in both informal financial communication and news-like textual environments.

2.5.4 Distribution of classes in the different datasets

Table 2.1. Distribution of classes in the different datasets

Dataset	Positive	Negative	Neutral	Total
Financial PhraseBank (50%)	1852	860	2128	4840
Financial PhraseBank (100%)	987	450	827	2264
Twitter Financial News	7740	1789	2398	11927
GOLD Headlines	4412	3814	376	8602

2.6 Proposed Approaches

2.6.1 Approach 1: Traditional sentiment analysis

This approach is based on domain specific modelling for classification of financial sentiment. It unifies feature extraction, fine-tuning, and classification into a single approach, based on 2 techniques as presented in Figure 2.4

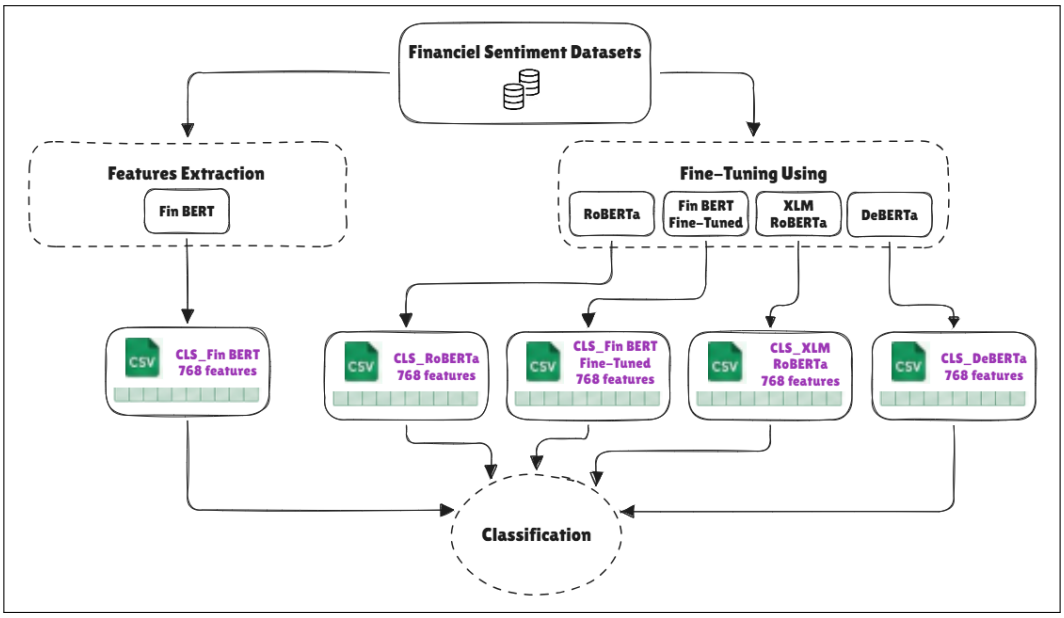


Figure 2.4. Traditional sentiment analysis

A. **Feature Extraction**

To enhancing financial sentiment classification via fine grained emotion modeling, the feature extraction stage transforms raw financial texts into numerical representations using a FinBERT model pre-trained on financial articles and datasets at the statement level. The contextual implications associated with the [CLS] code are extracted for each case, resulting in a 768-dimensional vector that embodies the semantic and sentiment features of the financial language, This representation is stored separately in a CSV file. providing a consistent and uniform input space for subsequent modeling stages.

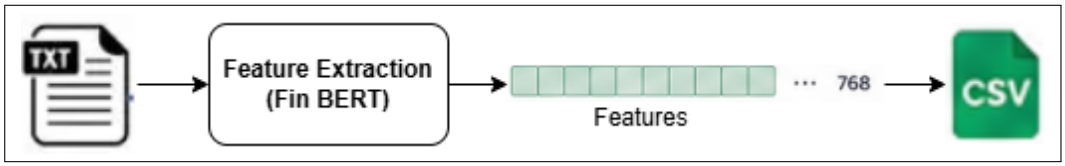


Figure 2.5. Method of Feature Extraction

B. **Fine-Tuning**

To improve financial sentiment classification, a fine-tuning process is applied in which XLM-RoBERTa, DeBERTa, RoBERTa, and FinBERT models are used on specific financial datasets at the phrase level. This helps create unique contextual representations for each model. Each structure goes through a process of being watched and adjusted, which helps it learn and remember the special meanings and feelings found in financial writings. For each case, the contextual embedding linked to the [CLS] code is taken from each model, creating four separate vectors that each have 768 dimensions. These embeddings are kept in four separate CSV files, which helps organize the features related to each model in a clear way, while also keeping the same format for use in later classification steps.

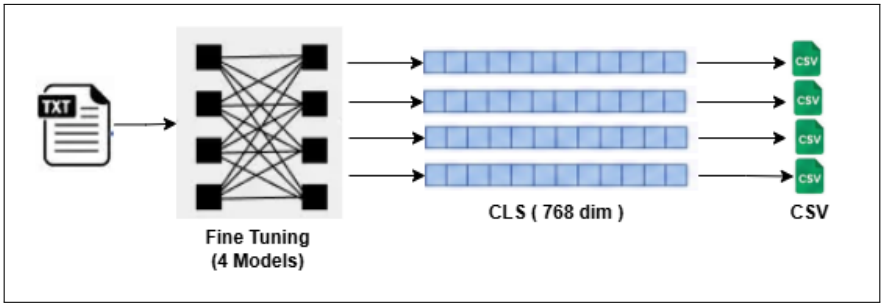


Figure 2.6. Method of Fine-Tuning

2.6.2 Approach 2: Fine Grained based sentiment analysis

In this approach, a pre-trained RoBERTa-based emotion model trained on the GoEmotions dataset is applied to capture fine-grained emotional polarity from text. The emotion vector extraction step uses a RoBERTa model, trained on the GoEmotions dataset, to analyze financial text and generate structured emotional representations. For each instance of input, the model takes the contextual embedding of the [CLS] token as input and predicts emotion scores across multiple categories. The outputs are stored separately in CSV files in two dimensions: a 768-dimensional vector and a 27-dimensional vector, for use in classification, as presented in Figure 2.7

A. **27-dimensional vector**

Fine-grained emotion probabilities across the 27 emotion categories in the GoEmotions taxonomy, enabling classification into one of the 27 emotion classes instead of an overall sentiment polarity.

B. **768-Dimensional vector**

It is the contextual semantic embedding obtained from the [CLS] token, which captures the global meaning and contextual information of financial text.

The choice of the 768-dimensional representation preserves compatibility with other transformer-based embeddings, such as RoBERTa and FinBERT outputs, thereby enabling consistent and efficient feature-level concatenation without introducing dimensional inconsistency.

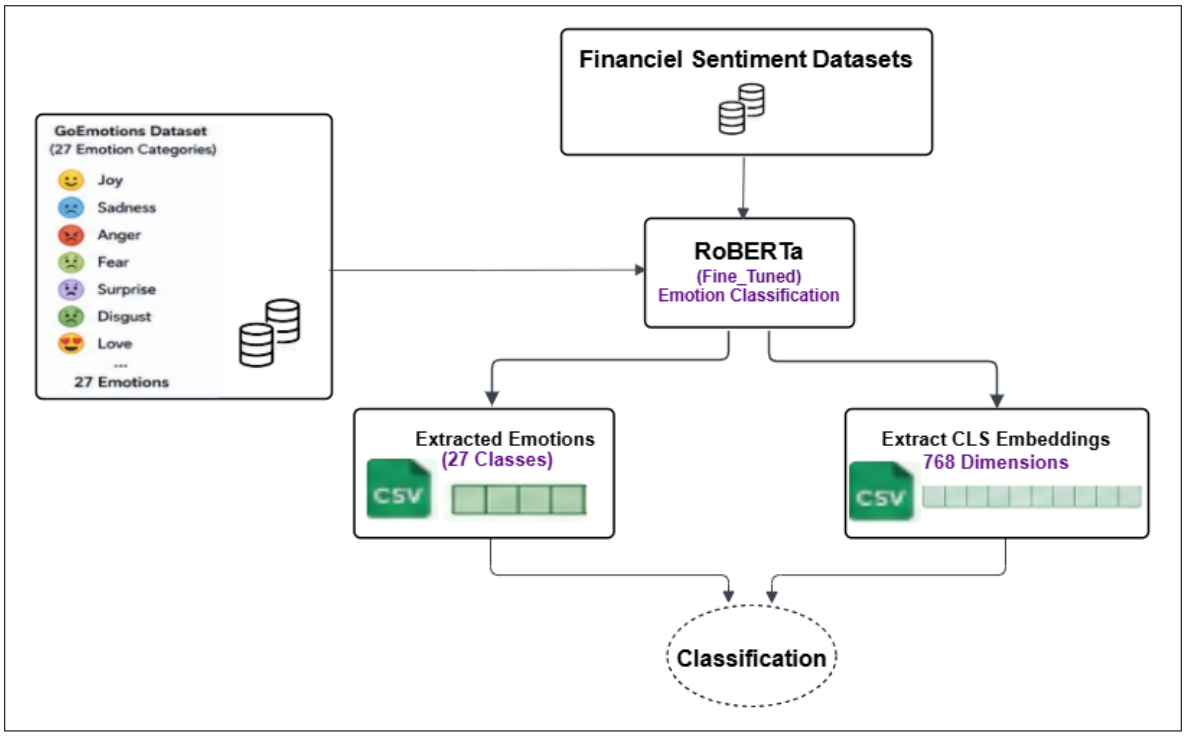


Figure 2.7. Fine Grained based sentiment analysis

2.6.3 Approach 3: Hybrid Feature Fusion

This approach combines features by concatenated the results from Approach 01,which uses financial embeddings, and Approach 02, which uses RoBERTa–GoEmotions. Four different ways of combining things are set up based on the schema, where each input instance is represented by aligned vectors originating from the same textual sample,as presented in Figure 2.8

A. **Concatenation 01:** combines the FinBERT feature extraction vector, 768 dimensions, with the RoBERTa [CLS] embedding, also 768 dimensions, to create a combined representation of 1536 dimensions in total.

B. **Concatenation 02:** combines embeddings from fine-tuned transformer models (FinBERT, XLM-RoBERTa, DeBERTa, RoBERTa) with the RoBERTa [CLS] embedding, resulting in a 1536-dimensional vector for each specific model pairing.

C. **Concatenation 03:** combines the FinBERT feature extraction embedding (768 dimensions) with the 27-dimensional predicted emotion vector, producing a final vector of 795 dimensions.

D. **Concatenation 04:** combines embeddings from fine-tuned transformer models with the 27-dimensional emotion vector, generating 795-dimensional vectors across different model configurations.

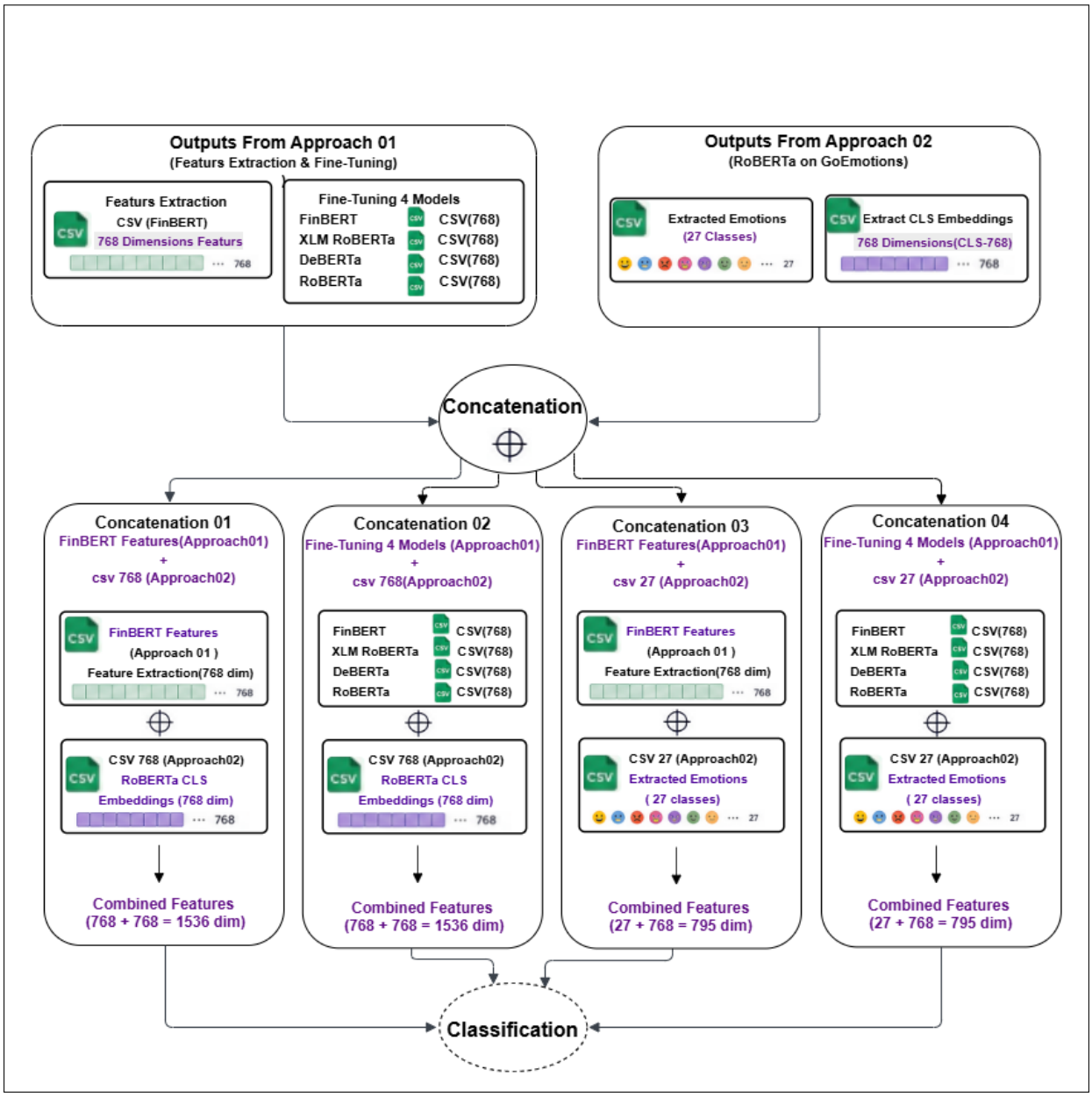


Figure 2.8. Hybrid Feature Fusion.

2.7 Classification setup

To evaluate the proposed framework for enhancing financial sentiment classification through fine-grained emotion modeling, three categories of models are employed: machine learning, deep learning, and attention-based architectures. This setup enables a consistent comparison of their ability to capture semantic and emotional information from financial text.

A. **Machine learning classifiers** like XGBoost, Support Vector Machines, Random Forest, Logistic Regression, CatBoost, and LightGBM are used on these fixed-length vectors that come from both pre-trained and fine-tuned embeddings.

B. **Deep learning classifiers** like CNN1D, LSTM, BiLSTM, GRU, BiGRU, transformers are used for classification in the space where words are represented using two different methods.

C. **Attention classifiers** like Self-Attention, Hierarchical Attention, Multi-Head Attention, and hybrid models such as BiGRU with Attention, are used for classification and to

emphasize important contextual information. This setup enables a consistent comparison and evaluation of attention-based methods within a unified framework.

The integration of these models allows a comprehensive evaluation of emotion-aware embeddings and highlights their effectiveness in improving financial sentiment classification.

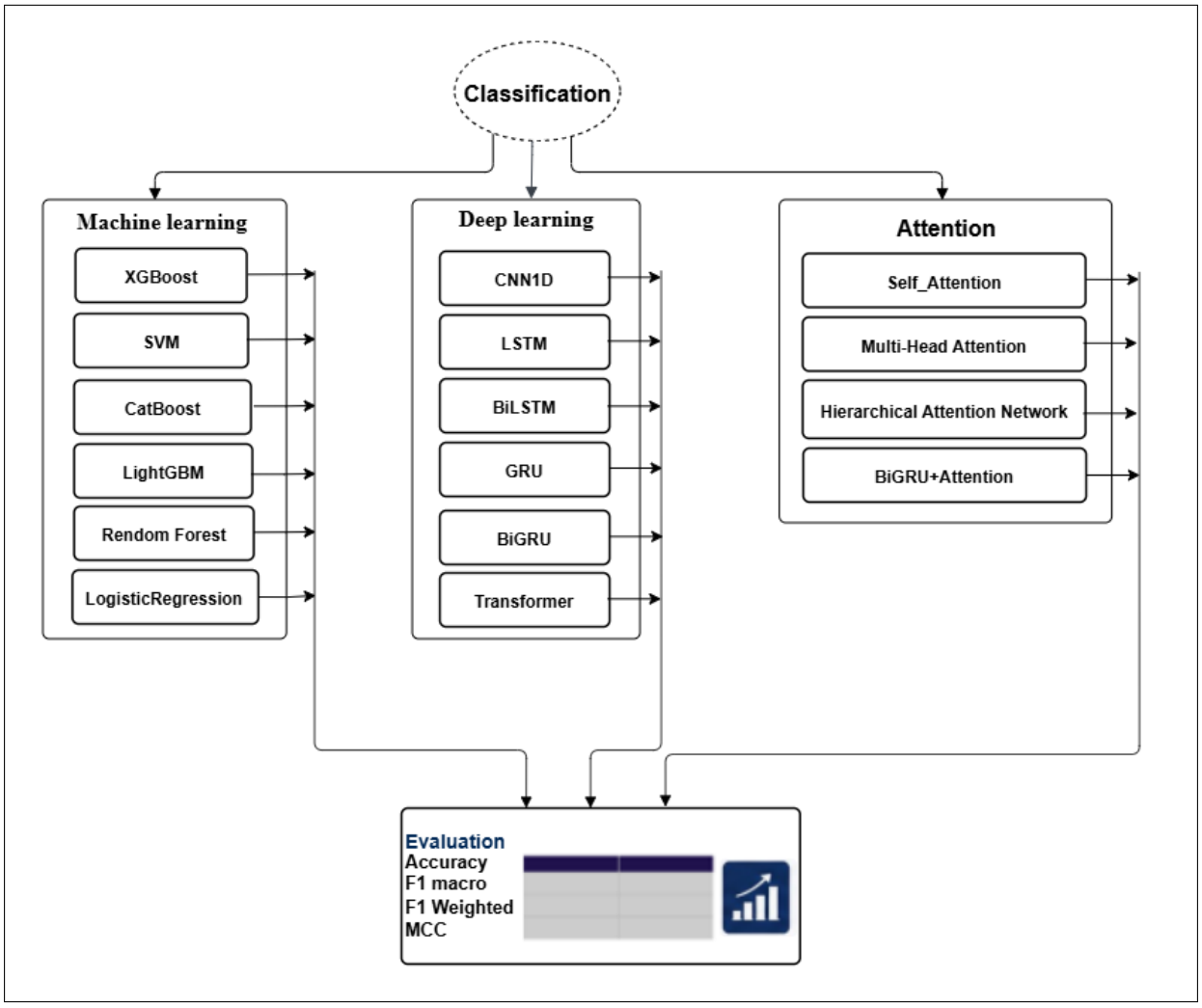


Figure 2.9. Classification Setup

2.8 Evaluation Metrics

2.8.1 Confusion matrix

The performance of the proposed models is evaluated using four metrics: Accuracy, F1-score Macro, F1-score Weighted, and Matthews Correlation Coefficient (MCC).These metrics are calculated from the confusion matrix and are widely used in financial sentiment analysis because they provide reliable evaluation under class imbalance conditions (Chicco & Jurman, 2020). Their use is particularly important in this project since the financial datasets contain imbalanced classes and fine-grained emotional variations extracted from the GoEmotions framework (Adams et al., 2023). Combining these evaluation metrics ensures a comprehensive and reliable analysis of machine learning, deep learning, and attention-based models . The confusion matrix is composed of four elements:

- **True Positive (TP):** correctly predicted positive samples.
- **True Negative (TN):** correctly predicted negative sample.

- **False Positive (FP):** negative samples incorrectly predicted as positive.
- **False Negative (FN):** positive samples incorrectly predicted as negative.

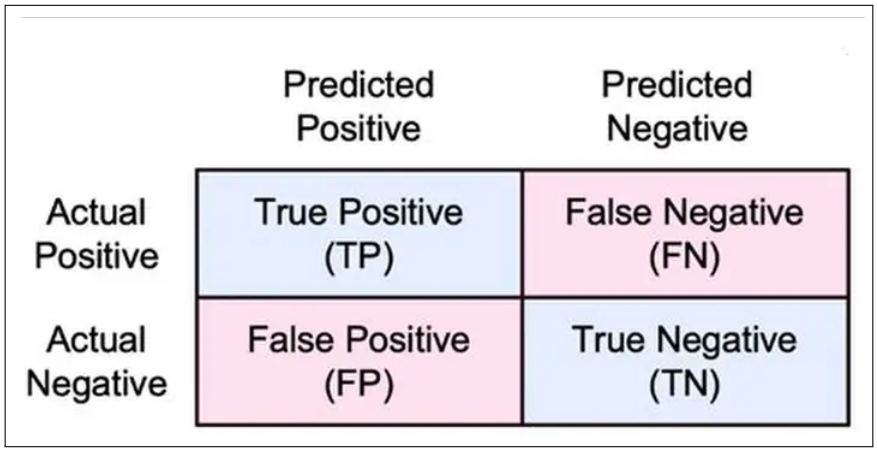


Figure 2.10. confusion matrix

2.8.2 Accuracy

The Accuracy represents the proportion of correct predictions across the entire test set. It is defined in formula 1:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

The importance of accuracy in this project is to provide an overall comparison between machine learning and deep learning models in integrating GoEmotions emotional representations into financial sentiment analysis. However, Accuracy can be misleading for imbalanced datasets, as high values can be achieved by favouring majority classes .

2.8.3 F1-score Macro

F1-score Macro computes the average F1-score independently for each class and then calculates their unweighted mean. Each class contributes equally regardless of its frequency.It is defined in formula 2:

$$\text{Macro F1 Score} = \frac{\sum_{i=1}^{n} \text{F1 Score}_i}{n} \tag{2.2}$$

F1-score Macro is important for fine-grained financial sentiment analysis because it evaluates whether the model performs well across all sentiment classes equally, including minority emotional categories such as fear, uncertainty, or pessimism (Grandini et al., 2020).

2.8.4 F1-score Weighted

F1-score Weighted calculates the average F1-score while considering the number of samples in each class .It is defined in formula 3:

$$\text{Weighted F1 Score} = \sum_{i=1}^{N} w_i \times \text{F1 Score}_i \tag{2.3}$$

The F1-score Weighted is important because financial datasets are naturally imbalanced. Weighted F1-score reflects the real contribution of dominant and minority classes simultaneously, making it more representative of practical model performance in financial sentiment classification tasks (Grandini et al., 2020).

2.8.5 Matthews Correlation Coefficient (MCC)

MCC is a correlation-based metric that evaluates the quality of binary classifications by considering all values of the confusion matrix simultaneously.It is defined in formula 4:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.4}$$

MCC is particularly valuable in this project because it provides a balanced evaluation even when the class distribution is highly unequal. Unlike Accuracy, MCC ensures that good performance is achieved on both majority and minority sentiment classes. Therefore, it is highly suitable for evaluating the effectiveness of GoEmotions-based fine-grained emotional representations in financial datasets.

2.9 conclusion

This chapter proposed the framework for 03 approaches for fine-grained financial sentiment analysis. Several financial datasets and the GoEmotions dataset were selected to capture both financial and emotional information. Three approaches were introduced: domain-specific modeling, emotion-driven modeling, and hybrid feature fusion combining financial and emotional vectors through different concatenation strategies.
The chapter also described the classification setup for machine learning, deep learning, and attention-based models, as well as the evaluation metrics used to assess performance under class imbalance conditions. Overall, this methodology provides a structured framework for studying the impact of fine-grained emotions on financial sentiment analysis and market trend prediction.

Chapter 3

Experiments and Results

3.1 Introduction

In this chapter, we evaluate the integration of fine-grained emotional information to improve financial sentiment analysis. Traditional models, relying only on domain-specific representations, often overlook subtle emotional cues affecting market behaviour. We therefore investigate three approaches: Approach 1 (Domain-Specific Modelling) using RoBERTa, FinBERT, DeBERTa, and XLM-RoBERTa for fine-tuning, along with FinBERT for feature extraction, Approach 2 (Emotion-Driven Modelling) based on CLS embeddings from a RoBERTa model pre-trained on GoEmotions, and Approach 3 (Hybrid Feature Fusion) combining financial and emotional representations fed into machine learning, deep learning, and attention-based models. The three approaches are benchmarked on PhraseBank (50% and 100%), Twitter Financial News, and Gold datasets using Accuracy, Macro F1, Weighted F1, and MCC metrics.

3.2 Development Environment

3.2.1 Hardware Specifications

A. **Kaggle** is an online platform that hosts data science competitions, datasets, and notebooks for data science projects. It allows data scientists and ma chine learning practitioners to collaborate, share knowledge, and participate in challenges that test their skills in various fields, including artificial intelligence, machine learning, and data analytics.
For this study, its access to robust GPUs like the NVIDIA P100 or dual T4 GPUs was critical for handling extensive training sessions.

B. **Google Colaboratory (Colab)** Colab provides a cloud-based development environment with access to free GPU acceleration (specifically the NVIDIA Tesla T4), which is highly suited for training deep learning models. It also supports real-time collaboration and allows the use of popular Python libraries for machine learning and data analysis

3.2.2 Software Environment & Libraries

A. **Python 3.12.13** is considered to be an interpreted programming language that is high-level and simple to read. It also supports many types of programming styles (e.g. procedural, object-oriented, functional, etc.). Python is heavily used in diverse fields such as web development, data science, artificial intelligence, and scientific computing, owing to its extensive standard library and active developer community. Furthermore, the versatility of the language and the comprehensive ecosystem of NLP libraries make Python a suitable choice for the implementation of this project (Pedregosa et al., 2011).

B. **PyTorch 2.11.0** is an open-source deep learning framework that is primarily known for its dynamic computational graph capabilities. PyTorch is the deep learning framework that will be utilized in this research for performing the tensor operations on the GPUs, fine-tuning the models, and implementing the fusion layer for the embeddings (Paszke et al., 2019).

C. **Scikit-Learn 1.6.1** an open-source Python machine learning library, is popular because of the vast number of tools and functions within it that allow for all aspects of data preprocessing, model selection, and all the most efficient implementations of machine learning algorithms including regression, classification, clustering, and dimensionality reduction. Scikit-learn works with NumPy and pandas so that you can experiment quickly and build reproducible and scalable machine learning pipelines for research or industry (Pedregosa et al., 2011).

D. **Pandas 2.2.2** is a fundamental Python library for working with structured (tabular) data. It offers highly optimized data structures, such as 'DataFrame' and 'Series', and supports operations like filtering, merging, aggregation, and time-series manipulation , all of which are essential for exploratory data analysis (EDA) and feature engineering (McKinney, 2010).

E. **NumPy 2.0.2** (Numerical Python) is the backbone for performing high-performance numerical operations in Python. It provides support for multi-dimensional arrays and a suite of mathematical functions, enabling efficient linear algebra computations, which are crucial in machine learning pipelines (Harris et al., 2020).

3.3 Experimental Setup

3.3.1 Data splitting

The datasets for this project are divided according to the classification approach used. For deep learning experiments (DL), the data is divided into 70% for training, 10% for validation, and 20% for testing to support model training, hyperparameter tuning, and performance evaluation. In machine learning experiments (ML), an 80% training and 20% testing split is used to maximize learning data and ensure reliable evaluation, as presented in Table 3.1.

These splitting strategies are important as they help in reducing overfitting, increasing model generalization and fair performance comparison between ML and DL approaches in fine-grained financial sentiment analysis using GoEmotions features.

Table 3.1. Dataset Splitting Strategy for ML and DL Experiments

Dataset	Split	Train	Validation	Test
ML Split (80% Train - 20% Test)				
Financial PhraseBank (50%)	80/20	3872	-	968
Financial PhraseBank (100%)	80/20	1811	-	453
Twitter Financial News	80/20	9542	-	2385
GOLD Headlines	80/20	6581	-	1645
DL Split (70% Train - 10% Validation - 20% Test)				
Financial PhraseBank (50%)	70/10/20	3388	484	968
Financial PhraseBank (100%)	70/10/20	1585	226	453
Twitter Financial News	70/10/20	8349	1193	2385
GOLD Headlines	70/10/20	5758	823	1645

3.3.2 Hyperparameters

This section presents the hyperparameter configurations used across the different categories of models evaluated in this study, namely classical machine learning, deep learning, and attention-based models.

A. **Machine learning**

The hyperparameters used to configure and tune the classical machine learning models are summarized in Table 3.2

Table 3.2. Hyperparameters Used for Classical ML Models

Model	**Hyperparameters**
XGBoost	n_estimators = 200, 500 max_depth = 4, 6 learning_rate = 0.05, 0.1
SVM	C = 1, 5, 10 kernel = rbf gamma = scale
Random Forest	n_estimators = 200, 300, 500 max_depth = 8, 10 max_features = default
Logistic Regression	C = 1, 10 solver = lbfgs penalty = l2 max_iter = 1000, 2000
CatBoost	iterations = 200, 400, 500 depth = 6 learning_rate = 0.05, 0.06, 0.1
LightGBM	num_leaves = 31, 63 learning_rate = 0.05, 0.1 n_estimators = 200, 300, 500

B. **Deep learning**

The hyperparameters used to configure and tune the deep learning models are summarized in Table 3.3

Table 3.3. Hyperparameters Used for Deep Learning Models

Model	**Hyperparameters**
GRU	hidden_units = 128 dropout = 0.3 learning_rate = 0.001
BiGRU	hidden_units = 128 dropout = 0.3 batch_size = 32
LSTM	hidden_units = 128 dropout = 0.3 learning_rate = 0.001
BiLSTM	hidden_units = 128 dropout = 0.3 sequence_length = 1
CNN-1D	filters = 64, 128, 256 kernel_size = 3, 5, 7 dropout = 0.3
Transformer	num_heads = 8 num_layers = 2 hidden_dim = 128

C. **Attention**

The hyperparameters used to configure and tune the attention-based models are summarized in Table 3.4

Table 3.4. Hyperparameters Configuration for Attention-Based Models

Model	Hyperparameters
Self-Attention	attention_dim = 768, 795, 1536 dropout = 0.3 num_heads = varied
Hierarchical Attention	attention_dim = 768, 795, 1536 hidden_units = 512, 256 dropout = 0.3
Multi-Head Attention	num_heads = 8 embed_dim = 768, 795, 1536 dropout = 0.3
BiGRU + Attention	hidden_units = 128 attention_dim = 256 dropout = 0.3

D. **Fine-tuning** The hyperparameters selected for the fine-tuning process were systematically applied across all evaluated architectures, including FinBERT, RoBERTa, DeBERTa-v3, and XLM-RoBERTa. For each of these models, the learning rate was tested at configurations of 2×10^{-5} and 1×10^{-5}, while the training was executed with a batch size of 32 over 5 epochs. The maximum sequence length was restricted to 128 and 256 tokens depending on the specific dataset requirements. Additionally, the AdamW optimizer was utilized to guide the gradient descent, combined with a dropout rate of 0.3 to prevent overfitting during the adaptation phase.

Table 3.5. Hyperparameters Used for Fine-Tuning Models

Model	Learning Rate	Batch Size	Epochs	Max Length	Optimizer	Dropout
Fine-tuning	2e-5/1e-5	32	5	128 / 256	AdamW	0.3

3.4 Experiment Results of Approach 1

This part presents the empirical results of Approach 1 for financial sentiment classification across four datasets (Phrasebank 50%, Phrasebank 100%, Twitter, and Gold). It evaluates two main strategies :feature extraction and fine-tuning, across three model families: ML, DL, and ATT mechanisms .

3.4.1 Feature Extraction

The performance of 16 different classifiers is examined using FinBERT via a feature extraction approach. The goal is to evaluate how effectively ML, DL and ATT models leverage these static, out-of-the-box financial embeddings.

Table 3.6. Approach 1 — Features Extraction: FinBERT classification results across four datasets

Model	Phrasebank 50%				Phrasebank 100%				Twitter				Gold			
	Acc	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*
Machine Learning (ML)																
SVM	89.28	88.10	89.30	80.81	98.66	97.88	98.67	97.53	**80.01**	**69.80**	**78.50**	**58.58**	**94.59**	**94.56**	**94.59**	**89.13**
RF	89.18	87.82	89.19	80.53	98.66	98.09	98.66	97.52	71.88	49.12	65.60	37.63	91.62	91.57	91.61	83.14
LR	87.84	87.38	87.87	78.22	98.44	97.56	98.44	97.12	76.28	66.73	75.72	52.45	91.37	91.33	91.37	82.65
XGB	**89.59**	**88.31**	**89.61**	**81.32**	98.88	98.09	98.89	97.94	76.91	64.34	74.92	51.50	92.28	92.24	92.28	84.48
CatB	89.28	87.95	89.31	80.82	**98.88**	**98.30**	**98.89**	**97.95**	75.78	62.01	73.37	48.62	92.04	92.00	92.04	83.99
LGBM	87.84	87.38	87.87	78.22	98.66	97.88	98.67	97.53	77.95	65.89	76.01	53.82	92.47	92.43	92.47	84.85
Deep Learning (DL)																
LSTM	88.76	88.00	88.77	79.84	**98.88**	**98.09**	**98.89**	**97.95**	80.01	71.96	79.54	60.01	91.59	91.56	91.60	83.13
BiLSTM	88.35	87.10	88.35	79.06	98.88	98.09	98.89	97.94	**81.18**	**73.66**	**80.80**	**62.50**	91.49	91.47	91.50	83.09
GRU	88.97	87.95	88.97	80.22	98.88	98.09	98.88	97.93	80.39	72.31	79.98	60.91	91.89	91.84	91.89	83.70
BiGRU	88.45	87.54	88.45	79.26	98.21	97.63	98.22	96.70	80.01	71.88	79.58	60.18	92.40	92.36	92.40	84.72
CNN-1D	89.07	87.96	89.07	80.23	98.66	98.30	98.65	97.53	71.96	56.61	69.79	41.03	90.58	90.56	90.59	81.33
Transf	**89.48**	**88.57**	**89.54**	**81.42**	98.88	98.09	98.88	97.93	74.77	67.67	75.43	53.75	92.00	91.94	91.99	83.91
Attention (ATT)																
Self-Att	**85.46**	**83.69**	**84.92**	**73.58**	98.21	97.21	98.21	96.72	79.63	71.57	79.29	59.58	**93.11**	**93.09**	**93.12**	**86.23**
Hier-Att	75.57	73.35	74.07	54.78	98.88	98.09	98.89	97.94	78.75	70.86	78.45	57.94	91.89	91.90	91.87	83.82
MH-Att	82.58	81.67	82.43	68.69	98.88	98.30	98.89	97.95	79.51	71.93	79.40	59.90	91.59	91.60	91.56	83.14
BiGRU+Att	88.35	87.25	88.34	78.95	**99.11**	**98.53**	**99.11**	**98.35**	**80.89**	**73.17**	**80.41**	**61.73**	92.40	92.38	92.41	84.81

Acc = Accuracy *F1-M* = F1 Macro *F1-W* = F1 Weighted *MCC* = Matthews Correlation Coefficient **bold** = best model in each category (ML/DL/ATT)

Model abbreviations: SVM = Support Vector Machine RF = Random Forest LR = Logistic Regression XGB = XGBoost CatB = CatBoost LGBM = LightGBM BiGRU+Att = BiGRU + Attention Transf = Transformer Self-Att = Self-Attention Hier-Att = Hierarchical Attention MH-Att = Multi-Head Attention

The analysis of Table 3.6 indicates that the classification of financial sentiment through feature extraction using FinBERT in Approach 1 yields high performance levels across the evaluated datasets. Among the three model families, BiGRU+Attention records the highest results overall. This model attains an F1-Weighted score of 99.11% and an MCC of 98.35% on the Phrasebank 100% dataset. It also achieves an F1-Weighted score of 80.41% and an MCC of 61.73% on the Twitter dataset. In the Deep Learning category, BiLSTM obtains the highest scores on the Twitter dataset. In the tarditional Machine Learning category, SVM achieves the highest scores on both the Twitter and Gold datasets.

In contrast, Hierarchical Attention, CNN-1D, and Random Forest produce the lowest results, particularly on Twitter dataset and Phrasebank 50%. The findings demonstrate clear variation in model performance according to data quality. Models reach near-saturation levels on the Phrasebank 100% and Gold datasets. The Twitter data functions as a discriminating benchmark that differentiates model capabilities. These outcomes illustrate the advantage of integrating bidirectional recurrent architectures with attention mechanisms for financial sentiment classification. They furthermore indicate that traditional classifiers maintain competitive performance when applied to high-quality contextual embeddings from FinBERT.

Figure 3.1 summarizes the MCC and F1-Weighted scores of the best-performing model from each category (ML, DL, and ATT) across all four datasets, further highlighting the consistent

superiority of Phrasebank 100% and the relatively weaker performance on the Twitter dataset.

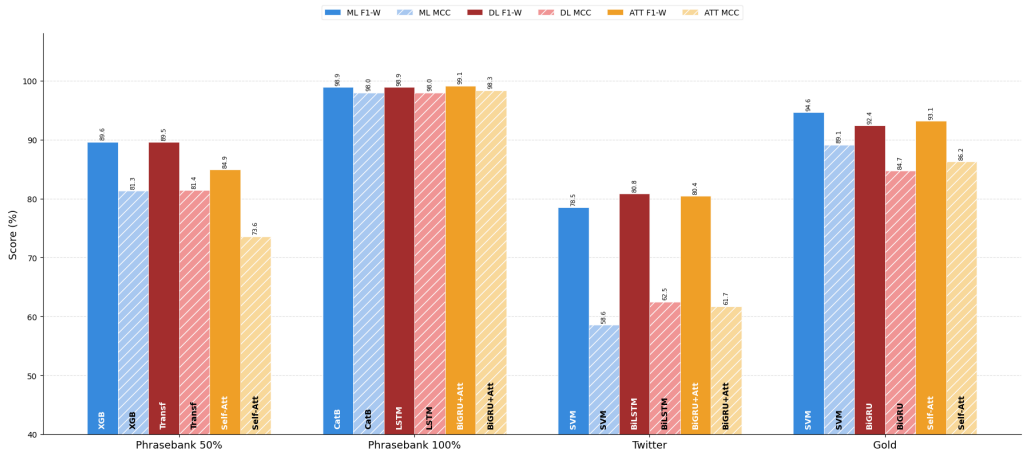


Figure 3.1. Best-performing model per category (ML, DL, ATT) — F1-Weighted and MCC scores across the four datasets (FinBERT feature extraction)

3.4.2 Fine-Tuning

The impact of fine-tuning four Transformer models (FinBERT, RoBERTa, DeBERTa, and XLM-RoBERTa) is evaluated in this section. The following results highlight the performance gains achieved when these models adapt directly to the specific nuances and vocabulary of each dataset.

Table 3.7. Approach 1 — Fine tuning : 4 Models classification results across four datasets

Model	Embedding	Phrasebank 50%				Phrasebank 100%				Twitter				Gold			
		Acc	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*
Machine Learning (ML)																	
SVM	FinBERT	86.49	86.49	85.14	75.88	98.21	97.47	98.21	96.71	92.29	89.23	92.22	84.88	96.11	96.07	96.10	92.22
	RoBERTa	**87.42**	**86.89**	**87.46**	**77.45**	99.11	98.73	99.10	98.35	94.84	93.16	94.76	89.92	96.35	96.32	96.35	92.69
	DeBERTa	84.12	82.12	84.21	71.78	99.11	98.73	99.10	98.35	91.83	89.14	91.66	83.83	95.62	95.58	95.61	92.90
	XLM-RoBERTa	85.36	83.85	85.49	74.58	96.88	96.18	96.85	94.23	95.81	94.10	95.83	91.96	95.74	95.70	95.73	91.51
RF	FinBERT	85.77	84.31	85.83	74.57	98.21	97.47	98.21	96.71	87.55	83.29	87.42	75.48	95.99	95.95	95.98	91.97
	RoBERTa	86.60	85.92	86.63	75.97	98.66	98.30	98.65	97.53	90.91	88.39	90.86	82.22	96.11	96.08	96.10	92.19
	DeBERTa	82.68	80.86	82.83	69.68	99.11	98.73	99.10	98.35	87.18	83.35	87.10	74.87	95.74	95.70	95.73	91.49
	XLM-RoBERTa	85.05	83.15	85.10	73.38	96.43	95.74	96.39	93.40	89.06	85.63	88.95	78.46	95.50	95.46	95.49	91.00
LR	FinBERT	86.19	85.14	86.25	75.46	99.11	98.73	99.10	98.35	85.96	81.42	85.94	72.69	95.74	95.71	95.74	91.45
	RoBERTa	86.29	85.25	86.31	75.44	98.66	98.30	98.65	97.53	89.82	87.08	89.84	80.33	95.62	95.59	95.62	91.21
	DeBERTa	82.37	80.58	82.52	69.13	98.66	98.30	98.65	97.53	83.53	78.74	83.63	68.46	94.77	94.72	94.76	89.51
	XLM-RoBERTa	85.57	84.02	85.69	74.87	95.54	94.48	95.49	91.78	88.22	84.85	88.27	77.36	95.62	95.58	95.61	91.25
XGB	FinBERT	86.29	84.86	86.35	75.52	98.66	97.89	98.66	97.54	96.56	94.78	96.54	93.30	96.35	96.32	96.35	92.69
	RoBERTa	86.80	86.17	86.82	76.31	99.11	99.15	99.10	98.35	98.16	97.44	98.15	96.42	96.35	96.33	96.35	92.67
	DeBERTa	83.30	81.25	83.38	70.26	99.11	98.73	99.10	98.35	97.36	96.17	97.33	94.86	95.99	95.95	95.98	91.96
	XLM-RoBERTa	85.46	83.77	85.47	74.00	95.98	94.89	95.93	92.60	97.57	96.32	97.57	95.30	95.74	95.70	95.73	91.49
CatB	FinBERT	86.19	84.88	86.25	75.31	98.66	98.31	98.66	97.53	96.48	94.65	96.45	93.13	**96.35**	**96.32**	**96.35**	**92.69**
	RoBERTa	87.42	86.94	87.45	77.45	98.66	98.30	98.65	97.53	**98.62**	**98.11**	**98.61**	**97.32**	96.35	96.32	96.35	92.69
	DeBERTa	83.30	81.32	83.39	70.30	99.11	98.73	99.10	98.35	97.57	96.46	97.54	95.27	95.74	95.71	95.73	91.48
	XLM-RoBERTa	85.15	83.23	85.16	73.44	96.88	96.18	96.85	94.23	97.57	96.34	97.56	95.28	95.50	95.46	95.49	90.99
LGBM	FinBERT	85.88	84.66	85.95	74.89	99.11	98.73	99.10	98.35	96.44	94.44	96.41	93.06	96.11	96.08	96.10	92.21
	RoBERTa	86.70	85.89	86.75	76.27	**99.55**	**99.58**	**99.55**	**99.17**	98.41	97.79	98.40	96.91	96.23	96.20	96.22	92.44
	DeBERTa	83.71	81.59	83.80	71.02	98.66	98.30	98.65	97.53	97.44	96.23	97.41	95.02	96.11	96.08	96.10	92.20
	XLM-RoBERTa	84.95	83.29	84.96	73.15	96.88	95.75	96.84	94.26	97.49	96.20	97.49	95.13	95.74	95.71	95.73	91.48
Deep Learning (DL)																	

Continued on next page...

Table 3.7 — *(continued)*

Model	Embedding	Phrasebank 50%				Phrasebank 100%				Twitter				Gold			
		Acc	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*
LSTM	FinBERT	86.80	85.37	86.83	76.31	97.32	96.23	97.33	95.09	93.63	90.77	93.62	87.60	96.47	96.45	96.47	92.92
	RoBERTa	86.91	86.23	86.91	76.50	98.66	98.30	98.65	97.53	96.40	95.32	96.36	92.97	96.47	96.45	96.47	92.90
	DeBERTa	83.51	82.09	83.62	70.86	98.66	98.31	98.66	97.53	95.26	93.67	95.22	90.73	95.74	95.70	95.73	91.53
	XLM-RoBERTa	84.95	83.08	85.02	73.26	95.54	94.48	95.49	91.78	96.81	95.36	96.81	93.82	95.74	95.70	95.73	91.51
BiLSTM	FinBERT	86.91	85.53	86.96	76.66	98.21	97.47	98.21	96.71	94.59	92.09	94.56	89.44	96.47	96.45	96.47	92.92
	RoBERTa	87.01	86.34	87.02	76.73	98.66	98.30	98.65	97.53	96.06	94.85	90.05	92.33	96.35	96.33	96.35	92.66
	DeBERTa	83.81	82.05	83.84	70.92	**99.11**	**98.73**	**99.10**	**98.35**	95.26	93.56	95.22	90.73	95.74	95.72	95.74	91.44
	XLM-RoBERTa	84.74	82.57	84.72	72.53	94.64	93.74	94.65	90.23	96.65	95.24	96.65	93.51	95.99	95.95	95.98	91.98
GRU	FinBERT	86.80	85.29	86.87	76.55	97.32	96.23	97.33	95.09	93.42	90.54	93.32	87.07	96.47	96.45	96.47	92.92
	RoBERTa	87.32	86.70	87.33	77.25	98.66	98.30	98.65	97.53	96.10	94.95	96.06	92.40	96.47	96.45	96.47	92.90
	DeBERTa	84.12	82.33	84.17	71.58	98.66	98.31	98.66	97.53	94.76	92.75	94.71	89.74	95.74	95.70	95.73	91.53
	XLM-RoBERTa	84.54	82.83	84.52	72.46	95.54	94.48	95.49	91.78	96.56	95.05	96.55	93.31	95.74	95.70	95.73	91.51
BiGRU	FinBERT	86.70	85.14	86.76	76.33	97.77	96.64	97.77	95.91	94.38	91.62	94.33	89.01	96.35	96.33	96.35	92.97
	RoBERTa	86.91	86.24	86.91	76.47	98.66	98.30	98.65	97.53	96.52	95.44	96.48	93.22	96.35	96.33	96.35	92.67
	DeBERTa	83.92	81.95	83.98	71.27	99.11	98.73	99.10	98.35	95.10	93.29	95.04	90.39	95.74	95.70	95.73	91.51
	XLM-RoBERTa	84.43	82.77	84.44	72.28	95.09	94.19	95.12	91.09	96.14	94.47	96.13	92.49	96.23	96.20	96.23	92.43
CNN-1D	FinBERT	**87.73**	**86.06**	**87.63**	**77.48**	95.09	93.67	95.14	91.07	96.23	94.22	96.17	92.63	96.47	96.45	96.47	92.90
	RoBERTa	87.22	86.39	87.27	77.28	97.77	96.82	97.81	95.95	**97.78**	**96.63**	**97.77**	**95.70**	96.23	96.20	96.23	92.42
	DeBERTa	81.96	80.17	82.09	68.17	98.21	97.44	98.24	96.74	96.56	95.09	96.55	93.33	95.86	95.84	95.86	91.68
	XLM-RoBERTa	84.33	82.29	84.18	71.86	96.88	95.91	96.85	94.21	96.69	95.13	96.01	93.59	95.74	95.71	95.73	91.48
Transf	FinBERT	84.74	83.32	84.91	73.77	98.21	97.90	98.22	96.71	90.86	87.48	91.17	83.50	**96.47**	**96.46**	**96.47**	**92.94**
	RoBERTa	86.70	85.79	86.73	76.30	98.66	98.31	98.66	97.53	97.32	95.84	97.32	94.81	95.99	95.95	95.98	91.98
	DeBERTa	84.02	81.36	83.92	70.82	99.11	98.73	99.10	98.35	89.73	87.20	90.35	82.43	96.11	96.08	96.10	92.18
	XLM-RoBERTa	85.05	83.21	84.26	73.18	95.09	94.19	95.12	91.09	94.05	91.96	93.90	88.43	96.11	96.08	96.10	92.20
Attention (ATT)																	
Self-Att	FinBERT	86.91	85.38	86.95	76.54	98.66	98.31	98.66	97.53	94.30	91.17	94.28	88.90	**96.59**	**96.57**	**96.59**	**93.15**
	RoBERTa	**87.63**	**86.82**	**87.62**	**77.68**	98.66	98.30	98.65	97.53	97.02	96.00	97.01	94.21	96.11	96.09	96.11	92.17
	DeBERTa	83.81	81.74	83.85	70.92	99.11	98.73	99.10	98.35	95.81	94.07	95.79	91.84	96.23	96.20	96.22	92.46
	XLM-RoBERTa	85.46	83.58	85.50	74.10	95.09	94.63	95.13	91.13	96.52	94.93	96.52	93.27	95.86	95.84	95.86	91.68
Hier-Att	FinBERT	87.63	86.32	87.67	77.85	98.66	98.31	98.66	97.53	93.71	90.53	93.65	87.80	96.47	96.45	96.47	92.93
	RoBERTa	87.42	86.81	87.42	77.34	99.11	98.73	99.10	98.35	97.02	96.15	97.01	94.20	96.23	96.20	96.23	92.43
	DeBERTa	84.23	82.42	84.27	71.75	99.11	98.73	99.10	98.35	95.60	93.85	95.58	91.42	95.50	95.45	95.49	91.06
	XLM-RoBERTa	85.88	84.23	85.96	75.07	95.09	94.63	95.13	91.13	96.35	94.69	96.36	92.95	96.23	96.20	96.22	92.46
MH-Att	FinBERT	86.70	85.02	86.76	76.27	97.32	96.23	97.33	95.09	94.64	91.99	94.60	89.53	96.47	96.45	96.47	92.90
	RoBERTa	87.01	86.37	87.01	76.65	98.66	98.30	98.65	97.53	**97.11**	**96.13**	**97.09**	**94.36**	96.11	96.08	96.10	92.20
	DeBERTa	84.43	82.82	84.47	72.00	**99.11**	**98.73**	**99.10**	**98.35**	95.52	93.67	95.51	91.31	95.99	95.96	95.98	91.94
	XLM-RoBERTa	85.26	83.46	85.29	73.79	95.09	94.63	95.12	91.09	96.94	95.51	96.93	94.05	96.23	96.20	96.22	92.45
BiGRU+Att	FinBERT	87.01	85.71	87.04	76.69	97.77	96.64	97.77	95.91	94.01	91.22	94.01	88.37	96.23	96.20	96.23	92.42
	RoBERTa	87.22	86.67	87.23	77.05	98.66	98.30	98.66	97.53	96.44	95.37	96.40	93.05	96.11	96.08	96.11	92.17
	DeBERTa	83.61	81.71	83.68	70.83	99.11	98.73	99.10	98.35	94.97	93.12	94.93	90.15	95.62	95.58	95.61	91.23
	XLM-RoBERTa	84.23	82.40	84.22	71.86	95.09	94.19	95.12	91.09	96.73	95.30	96.72	93.65	96.11	96.08	96.10	92.20

Acc = Accuracy *F1-M* = F1 Macro *F1-W* = F1 Weighted *MCC* = Matthews Correlation Coefficient **bold** = best per section

Model abbreviations: SVM = Support Vector Machine RF = Random Forest LR = Logistic Regression XGB = XGBoost CatB = CatBoost LGBM = LightGBM BiGRU+Att = BiGRU + Attention Transf = Transformer Self-Att = Self-Attention Hier-Att = Hierarchical Attention MH-Att = Multi-Head Attention

The analysis of Table 3.7 presents the classification performance obtained through fine-tuning of various models across four embeddings (FinBERT, RoBERTa, DeBERTa, and XLM-RoBERTa) on four financial sentiment datasets. The results indicate variation in model performance according to both architecture category and embedding type.

Among all configurations, the combination of RoBERTa with CatBoost yields the highest results on the Twitter dataset (F1-W = 98.61%, MCC = 97.32%). On the Phrasebank 100% dataset, RoBERTa with LGBM records the highest scores (F1-W = 99.55%, MCC = 99.17%).

In the Machine Learning category, CatBoost and LGBM paired with RoBERTa achieve the strongest performance on the more challenging Twitter dataset. In the Deep Learning category, CNN-1D with RoBERTa attains the highest scores on Twitter (F1-W = 97.77%, MCC = 95.70%). In the Attention category, MH-Att with RoBERTa obtains the highest Twitter scores (F1-W = 97.09%, MCC = 94.36%), while Self-Att with FinBERT reaches the highest MCC on the Gold dataset (MCC = 93.15%).
RoBERTa consistently outperforms FinBERT, DeBERTa, and XLM-RoBERTa across most model families, particularly on the Twitter dataset. In contrast, DeBERTa and XLM-RoBERTa yield lower performance levels on several configurations. The findings show that model performance varies substantially according to dataset characteristics: near-saturation levels appear on Phrasebank 100% and Gold, whereas the Twitter dataset provides greater differentiation between models and embeddings.
These results demonstrate that fine-tuning improves performance compared to feature extraction embeddings, particularly when RoBERTa is used as the base model. Moreover, gradient boosting models (CatBoost and LGBM) remain competitively effective in the fine-tuning setting for financial sentiment classification.
Figure 3.2 illustrates the best F1-Weighted and MCC scores per embedding across all four datasets for each model category (ML, DL, and ATT), further confirming the dominance of RoBERTa under the fine-tuning strategy.

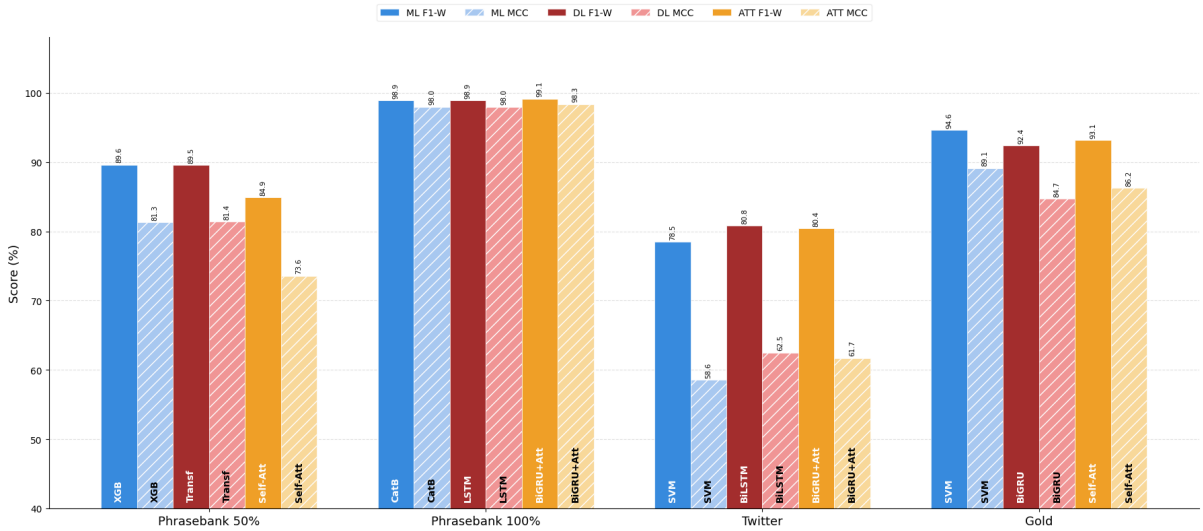


Figure 3.2. Approach 1 — Fine-tuning: best F1-Weighted and MCC scores per embedding (ML / DL / ATT) across four datasets

3.4.3 Discussion Results of Approach 1

The comparison between the feature extraction and fine-tuning approaches in Approach 1 shows differences in performance across the four datasets.

- F1-Weighted and MCC are selected as the primary evaluation metrics because they offer greater robustness to class imbalance, which is particularly evident in the Twitter dataset.
- The Phrasebank datasets record the highest classification performance overall. Phrasebank 100% outperforms Phrasebank 50% because it contains only sentences with full annotator agreement, thereby reducing label ambiguity. Phrasebank 50% relies on majority agreement and introduces greater label uncertainty.
- On Phrasebank 100%, the best F1-Weighted score reached 99.11% in feature extraction and 99.55% in fine-tuning, suggesting that models already achieve near-saturation

performance on highly reliable data with feature extraction. On Phrasebank 50%, the best F1-Weighted score was 89.59% in feature extraction compared to 87.73% in fine-tuning.

- The Gold dataset produced strong and stable results in both strategies. The best F1-Weighted score increased from 94.59% in feature extraction to 96.59% in fine-tuning , while the MCC improved from 89.13% to 93.15%.
- In contrast, the Twitter dataset recorded the lowest performance across all models. Its unstructured social media nature , characterized by informal language, slang, abbreviations, and noise, makes it particularly challenging. Under feature extraction, the best model achieved 80.41% F1-Weighted and 61.73% MCC. Fine-tuning led to a major improvement, reaching 98.61% F1-Weighted and 97.32% MCC. These outcomes reflect the high annotation quality and formal language of the dataset.
- Regarding the two strategies, the feature extraction approach delivered competitive but more limited results, with BiGRU+Attention performing best on the difficult Twitter dataset. The fine-tuning approach, however, produced markedly superior outcomes across all datasets, particularly on the Twitter dataset. RoBERTa combined with gradient boosting models (CatBoost and LGBM) achieved the highest scores overall.
- These findings indicate that fine-tuning enables better adaptation to noisy data than feature extraction. Dataset formality and annotation quality constitute key determinants of classification performance.

3.5 Experiment Results of Approach 2

This part presents the empirical results of Approach 2, focusing on financial sentiment classification using RoBERTa embeddings pre-trained on GoEmotions across four datasets. The primary objective here is to evaluate performance variations across three model families ML,DL, and ATT mechanisms .while directly comparing the impact of low-dimensional text representations (the 27-dimensional emotion classification layer output from the RoBERTa model) versus full-dimensional representations (the 768-dimensional CLS token output).presented in table 3.8:

Table 3.8. Approach 2 — RoBERTa Embeddings: classification results across four datasets

Model	Dim	Phrasebank 50%				Phrasebank 100%				Twitter				Gold			
		Acc	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*
Machine Learning (ML)																	
SVM	27	78.04	75.97	78.09	61.10	91.74	89.91	91.70	84.91	67.48	61.80	68.12	41.03	81.29	81.19	81.29	62.38
	768	**83.81**	**82.25**	**83.84**	**70.97**	**97.32**	**96.71**	**97.31**	**95.04**	**82.77**	**77.79**	**82.87**	**66.98**	**94.23**	**94.19**	**94.22**	**88.39**
RF	27	79.79	78.19	79.78	63.74	93.75	92.78	93.63	88.44	72.80	66.81	73.00	48.58	86.94	86.86	86.93	73.72
	768	81.24	79.22	81.09	65.97	95.31	94.49	95.23	91.32	79.30	72.66	78.86	58.58	92.22	92.17	92.21	84.36
LR	27	73.51	71.35	73.35	52.17	88.39	87.42	88.63	79.90	64.33	58.50	65.13	35.56	63.55	61.65	62.27	26.34
	768	80.10	78.65	80.34	66.08	97.10	96.62	97.09	94.64	77.20	72.45	77.78	58.78	91.74	91.68	91.73	83.38
XGB	27	80.82	78.83	80.55	64.77	93.53	92.50	93.37	88.03	**77.95**	**70.02**	**76.98**	**54.69**	86.27	86.22	86.28	72.49
	768	82.37	80.42	82.19	67.73	96.43	95.76	96.40	93.40	82.56	75.93	81.87	64.61	92.89	92.86	92.89	85.72
CatB	27	**80.82**	**78.50**	**80.53**	**64.99**	92.19	91.50	92.19	85.78	77.12	68.51	75.93	52.68	**87.30**	**87.26**	**87.31**	**74.55**
	768	81.96	79.70	81.75	67.12	96.21	95.03	96.18	93.03	81.60	74.47	80.75	62.48	92.53	92.48	92.53	84.97
LGBM	27	81.13	79.57	81.16	66.20	**94.64**	**94.01**	**94.52**	**90.10**	75.78	69.41	75.80	53.18	86.63	86.59	86.65	73.21

Continued on next page...

Table 3.8 — *(continued)*

Model	Dim	Phrasebank 50%				Phrasebank 100%				Twitter				Gold			
		Acc	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*
	768	83.51	81.67	83.40	69.97	96.21	95.49	96.14	93.00	82.48	76.47	82.10	64.98	93.20	93.16	93.20	86.32
Deep Learning (DL)																	
LSTM	27	79.38	76.79	78.84	61.91	89.29	87.44	89.17	80.20	72.30	60.74	70.30	41.38	76.49	76.48	76.51	53.45
	768	83.20	82.04	81.78	69.48	**98.44**	**98.32**	**98.44**	**97.12**	82.65	76.40	82.27	65.31	**93.80**	**93.77**	**93.80**	**87.54**
BiLSTM	27	79.28	76.52	78.69	61.74	**90.18**	**88.18**	**89.99**	**81.83**	73.01	61.52	70.96	42.80	82.44	82.39	82.46	64.83
	768	**84.12**	**82.52**	**73.16**	**71.16**	97.10	96.31	97.11	94.66	**82.69**	**77.14**	**82.45**	**65.65**	93.62	93.59	93.62	87.19
GRU	27	79.59	76.94	78.98	62.25	90.18	88.24	89.91	81.77	71.92	60.46	69.96	40.65	78.68	78.64	78.70	57.46
	768	83.51	82.20	80.24	70.24	97.77	97.35	97.76	95.88	82.61	76.55	82.20	65.12	93.74	93.70	93.74	87.41
BiGRU	27	79.90	77.19	79.27	62.83	90.18	88.16	89.96	81.88	73.22	62.53	71.50	43.65	82.44	82.42	82.47	65.03
	768	82.68	81.29	78.66	68.66	97.77	97.47	97.76	95.87	81.94	76.14	81.67	64.14	93.86	93.83	93.87	87.67
CNN-1D	27	**80.10**	**78.02**	**79.77**	**63.52**	90.18	88.51	90.05	81.93	**75.48**	**66.56**	**74.37**	**49.51**	**83.96**	**83.88**	**83.96**	**67.75**
	768	82.78	80.02	82.58	68.42	97.10	96.27	97.08	94.66	81.01	73.29	79.87	60.96	91.25	91.23	91.26	82.55
Transf	27	77.84	75.03	77.48	59.56	90.07	89.24	90.77	83.48	71.75	62.06	70.56	41.65	83.11	83.09	83.13	66.38
	768	78.87	61.21	78.57	75.58	97.54	97.03	97.55	95.50	76.91	70.31	76.90	55.33	92.83	92.77	92.82	85.62
Attention (ATT)																	
Self-Att	27	79.69	77.05	79.20	62.44	**91.74**	**90.00**	**91.64**	**84.81**	73.72	64.63	72.67	46.07	**85.24**	**85.19**	**85.25**	**70.45**
	768	83.71	81.52	83.64	70.48	**97.77**	**97.57**	**97.77**	**99.79**	**83.11**	**77.45**	**82.73**	**66.11**	93.13	93.10	93.14	86.21
Hier-Att	27	78.04	73.72	76.22	59.67	90.85	89.03	90.66	83.11	75.23	66.00	74.08	48.72	75.21	75.21	75.22	51.07
	768	82.89	81.60	82.85	68.93	97.54	96.95	97.55	99.77	82.48	76.59	82.14	65.05	92.89	92.85	92.89	85.70
MH-Att	27	**80.10**	**77.84**	**79.87**	**63.74**	91.74	90.12	91.48	84.72	**75.19**	**67.09**	**74.40**	**49.71**	82.38	82.37	82.40	65.03
	768	82.06	80.90	82.16	67.82	96.21	95.43	96.26	99.21	82.27	76.99	82.24	65.52	93.26	93.20	93.25	86.45
BiGRU+Att	27	78.87	75.95	78.11	60.89	90.18	88.22	89.88	81.83	72.72	61.24	70.66	42.23	82.44	82.41	82.46	64.94
	768	**84.12**	**82.74**	**83.24**	**71.34**	97.54	97.14	97.54	95.47	82.27	76.91	82.16	65.25	**93.92**	**93.89**	**93.92**	**87.78**

Acc = Accuracy *F1-M* = F1 Macro *F1-W* = F1 Weighted *MCC* = Matthews Correlation Coefficient (%) **bold** = best per section & dimension

Model abbreviations: SVM = Support Vector Machine RF = Random Forest LR = Logistic Regression XGB = XGBoost CatB = CatBoost LGBM = LightGBM BiGRU+Att = BiGRU + Attention Transf = Transformer Self-Att = Self-Attention Hier-Att = Hierarchical Attention MH-Att = Multi-Head Attention

The analysis of Table 3.8 (Approach 2 — RoBERTa Embeddings) reveals substantial differences in classification performance between the 27-dimensional and 768-dimensional embeddings across the three model categories.

In the **Machine Learning category** SVM with 768 dimensions records the best overall results, ranking first among all ML models across datasets. On Phrasebank 100%, LGBM records the highest F1-Weighted score of 94.52% at 27 dimensions, which increases to SVM leading at 768 dimensions with 97.31%. On Phrasebank 50%, the best performance at 27 dimensions is achieved by CatBoost with an F1-Weighted score of 80.53%, while at 768 dimensions, SVM reaches 83.84%. For the Gold dataset, the highest MCC at 27 dimensions is recorded by CatBoost at 74.55%, which shifts to SVM at 768 dimensions with an MCC of 88.39%. On the Twitter dataset, XGBoost achieves the highest F1-Weighted score of 76.98% at 27 dimensions, whereas SVM leads at 768 dimensions with 82.87%. Conversely, Logistic Regression (LR) emerges as the weakest model in this category, particularly on the Gold dataset where it drops to an MCC of 26.34% at 27 dimensions, and remains the lowest at 768 dimensions with an MCC of 83.38%.

In the **Deep Learning category** BiLSTM with 768 dimensions and CNN-1D with 27

dimensions emerge as the top performers within their respective dimensional tiers. On Phrasebank 100%, CNN-1D captures the top spot at 27 dimensions with an F1-Weighted score of 90.05%, while LSTM achieves the highest score at 768 dimensions with an F1-Weighted score of 98.44%. On Phrasebank 50%, CNN-1D yields the highest F1-Weighted score at 27 dimensions with 79.77%, while BiLSTM leads at 768 dimensions with 73.16%. On the Gold dataset, CNN-1D leads at 27 dimensions with an MCC of 67.75%, whereas BiLSTM captures the top spot at 768 dimensions within the pure DL family with an MCC of 87.19%. On the Twitter dataset, CNN-1D obtains the top MCC at 27 dimensions with 49.51%, while BiLSTM secures the highest MCC at 768 dimensions with 65.65%. In contrast, the standard Transformer model yields the lowest results in this family, scoring an F1-Weighted of 77.48% on Phrasebank 50% at 27 dimensions, and registering the lowest MCC on Twitter at 768 dimensions with 55.33%.

In the **Attention category** BiGRU+Attention (768 dim) and Self-Attention (768 dim) deliver the strongest results. On Phrasebank 100%, Self-Attention leads at 27 dimensions with an F1-Weighted score of 91.64%, while at 768 dimensions, Self-Attention remains the top model with an exceptional MCC of 99.79%. On Phrasebank 50%, MH-Att achieves the best F1-Weighted score of 79.87% at 27 dimensions, while BiGRU+Attention leads at 768 dimensions with 83.24%. On the Gold dataset, Self-Attention leads at 27 dimensions with an MCC of 70.45%, while BiGRU+Attention yields the highest MCC at 768 dimensions with 87.78%. On the Twitter dataset, MH-Att secures the highest F1-Weighted score of 74.40% and an MCC of 49.71% at 27 dimensions, whereas Self-Attention outperforms the others at 768 dimensions with an F1-Weighted score of 82.73% and an MCC of 66.11%. On the other end of the spectrum, Hierarchical Attention (Hier-Att) demonstrates the poorest adaptation, trailing with an F1-Weighted score of 76.22% on Phrasebank 50% at 27 dimensions, and finishing last on the Gold dataset at 768 dimensions with an MCC of 85.70%.

Overall, the 768-dimensional embeddings consistently outperform the 27-dimensional ones in all three categories (ML, DL, and ATT), with larger gains observed on the more challenging Twitter and Gold datasets. These results confirm the advantage of using the full-dimensional RoBERTa representations for financial sentiment classification.
Figure 3.18 presents the F1-Weighted and MCC scores of the best model per category (ML,DL ,ATT) at 768 dimensions.

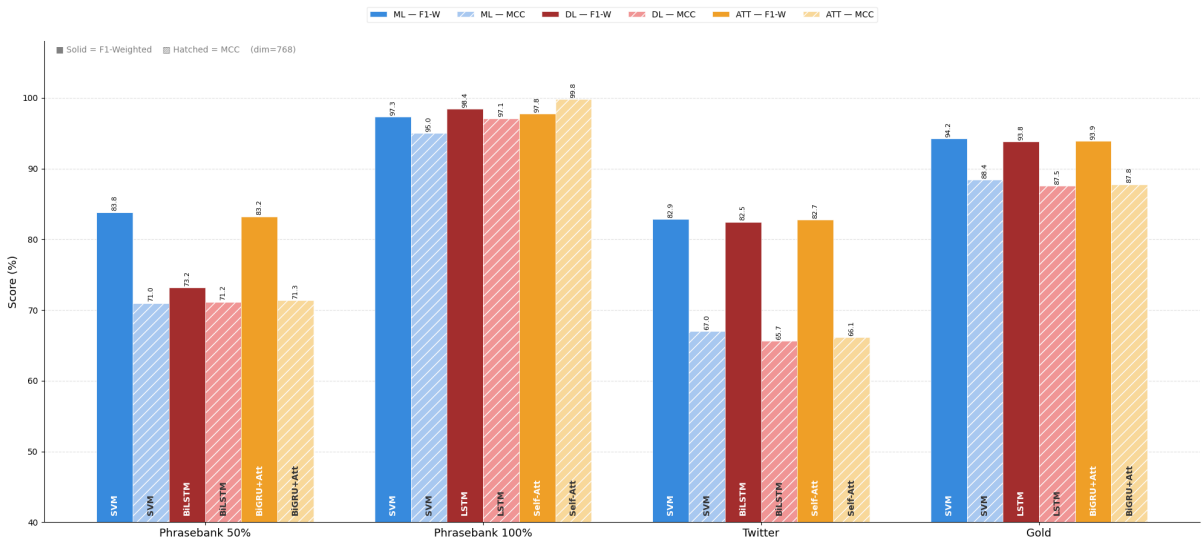


Figure 3.3. Approach 2 — RoBERTa Feature Extraction (Dim=768): F1-Weighted and MCC scores of the best-performing model per category (ML, DL, ATT) across four datasets

3.5.1 Discussion Results of Approach 2

The results of Approach 2 highlight three primary determinants of classification performance: embedding dimensionality, dataset annotation quality, and architectural design.

- The 768-dimensional [CLS] representation systematically outperforms the 27-dimensional emotion-probability vector across all models, families, and datasets, with substantial MCC gains on Twitter and Gold datasets. This confirms that the 27-dimensional emotion-probability vector, derived from a GoEmotions fine-tuned projection, loses critical financial sentiment cues that are preserved in the full 768-dimensional [CLS] representation, making the latter a substantially more suitable feature space for financial sentiment classification.
- Dataset characteristics directly condition performance. Phrasebank 100% achieves near-perfect scores due to full annotator agreement, while Phrasebank 50% introduces label uncertainty through majority-only agreement, producing higher variance. Gold yields strong and stable results owing to its formal language and annotation quality, with the best F1-Weighted reaching 93.92% at 768 dimensions (BiGRU+Att). Twitter records the lowest performance across all settings, with a persistent ceiling of 83.11% F1-Weighted and 66.11% MCC at 768 dimensions, attributable to the mismatch between RoBERTa's formal pretraining corpus and informal social media text.
- At 768 dimensions, SVM leads the ML category, BiLSTM leads the DL category, and BiGRU+Att leads the Attention category. At 27 dimensions, CNN-1D, MH-Att, and CatBoost achieve the relatively strongest results within their families, though all remain substantially below their 768-dimensional counterparts.
- Overall, the feature extraction paradigm delivers competitive results on clean, domain-aligned datasets such as Phrasebank 100% and Gold, but reaches its practical limits on informal and noisy benchmarks such as Twitter, where the distributional mismatch between RoBERTa's formal pretraining corpus and social media text cannot be compensated by classifier or dimensionality optimization alone.

3.6 Experiment results of Approach 3

3.6.1 Concatenation FinBERT Features Extraction with Roberta - GoEmotion (27-768)

This part presents the results of Concatenation 01:Approche 3 concatenation FinBERT features extraction for financial sentiment classification using concatenated feature representations across four datasets. Two fusion strategies, namely 786+27 and 768+768, are evaluated using ML, DL, and ATT models. The corresponding results are summarized in Table 3.9.

Table 3.9. Approach 3 — Concatenation Feature extraction: classification results across four datasets

Model	Concat.	Phrasebank 50%				Phrasebank 100%				Twitter				Gold			
		Acc	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*
Machine Learning (ML)																	
SVM	768+27	89.46	88.73	89.47	81.03	**99.11**	**98.86**	**99.10**	**98.35**	**82.59**	**76.28**	**82.41**	**65.72**	**93.11**	**93.09**	**93.12**	**86.21**
	768+768	**90.39**	**89.59**	**90.40**	**82.71**	98.81	98.42	98.81	97.80	**85.03**	**80.16**	**84.92**	**70.62**	**94.93**	**94.91**	**94.93**	**89.82**

Continued on next page...

Table 3.9 — *(continued)*

Model	Concat.	Phrasebank 50%				Phrasebank 100%				Twitter				Gold			
		Acc	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*
RF	768+27	89.87	89.01	89.87	81.79	**99.11**	**98.86**	**99.10**	**98.35**	78.49	70.59	77.51	55.97	90.17	90.13	90.18	80.26
	768+768	89.77	88.97	89.76	81.57	**99.11**	**98.86**	**99.10**	**98.35**	80.79	73.65	80.00	60.89	92.91	92.88	92.91	85.78
LR	768+27	86.88	85.93	86.91	76.48	98.51	97.85	98.53	97.27	73.83	67.63	74.75	53.48	90.37	90.35	90.39	80.80
	768+768	87.80	87.06	87.83	77.48	98.81	98.42	98.81	97.80	77.69	72.96	78.28	59.86	93.62	93.60	93.62	87.27
XGB	768+27	**90.28**	**89.63**	**90.28**	**82.49**	**99.11**	**98.86**	**99.10**	**98.35**	81.97	74.78	81.13	63.17	91.79	91.76	91.80	83.55
	768+768	89.25	88.21	89.23	80.55	**99.11**	**98.86**	**99.10**	**98.35**	83.35	76.80	82.62	65.40	94.12	94.10	94.13	88.20
CatB	768+27	89.56	88.77	89.54	81.18	98.81	98.42	98.81	97.80	77.52	68.67	76.19	53.26	91.69	91.66	91.70	83.33
	768+768	89.25	88.28	89.22	80.52	**99.11**	**98.71**	**99.11**	**98.35**	82.97	76.72	82.28	65.40	93.11	93.09	93.12	86.20
LGBM	768+27	89.56	88.72	89.55	81.28	98.81	98.42	98.81	97.80	82.30	75.51	81.56	63.96	91.39	91.35	91.39	82.72
	768+768	89.66	88.73	89.64	81.28	98.81	98.42	98.81	97.80	84.06	77.99	83.43	67.71	93.01	92.98	93.01	86.00
Deep Learning (DL)																	
LSTM	768+27	85.87	80.73	87.47	64.61	98.51	98.15	98.52	97.26	81.24	73.93	84.33	58.53	**92.91**	**92.89**	**92.91**	**85.84**
	768+768	**81.83**	**72.15**	**80.54**	**65.52**	98.51	98.00	98.52	97.29	**81.55**	**69.52**	**84.93**	**65.97**	93.31	93.29	93.32	86.61
BiLSTM	768+27	82.73	81.14	85.90	71.79	98.51	98.00	98.52	97.26	**78.95**	**68.56**	**77.16**	**64.45**	90.48	90.44	90.48	80.91
	768+768	80.94	69.57	82.33	59.03	98.51	98.15	98.52	97.26	82.27	75.98	78.35	67.37	93.31	93.29	93.32	86.63
GRU	768+27	84.78	80.02	82.90	66.30	**99.11**	**98.71**	**99.11**	**98.36**	83.16	73.74	83.24	59.10	92.20	92.17	92.20	84.35
	768+768	79.39	71.59	81.25	64.40	98.21	97.86	98.21	96.70	82.63	76.51	83.29	56.21	93.01	92.99	93.02	86.09
BiGRU	768+27	**87.61**	**81.62**	**85.03**	**73.02**	98.81	98.43	98.82	97.82	79.67	75.81	77.66	64.33	92.30	92.29	92.31	84.72
	768+768	80.05	74.01	79.43	65.46	98.81	98.58	98.80	97.79	85.51	70.13	84.75	59.10	**94.33**	**94.30**	**94.33**	**88.61**
CNN-1D	768+27	84.75	79.30	80.02	63.97	97.02	96.43	97.03	94.51	78.20	68.27	81.67	57.31	91.79	91.76	91.80	83.52
	768+768	84.57	69.18	78.00	58.14	**98.81**	**98.43**	**98.81**	**97.80**	80.16	71.67	78.94	66.83	92.50	92.47	92.50	84.94
Transf	768+27	81.31	74.51	82.33	62.87	98.81	98.29	98.80	97.78	78.74	72.30	78.25	63.64	91.89	91.84	91.89	83.70
	768+768	79.08	73.53	82.75	65.37	97.62	96.71	97.63	95.61	80.84	74.44	82.73	64.31	92.60	92.59	92.61	85.37
Attention (ATT)																	
Self-Att	768+27	86.61	77.10	82.37	68.45	98.51	98.00	98.52	97.29	**85.89**	**68.52**	**84.08**	**66.40**	92.71	92.71	92.69	85.45
	768+768	**88.22**	**77.54**	**85.77**	**68.71**	98.21	97.58	98.23	96.76	**85.93**	**77.37**	**79.77**	**66.76**	**95.20**	**95.18**	**95.20**	**90.35**
Hier-Att	768+27	**88.79**	**82.19**	**82.21**	**77.06**	97.92	97.42	97.92	96.15	83.81	71.23	84.83	56.07	**93.01**	**93.01**	**92.98**	**85.95**
	768+768	83.04	88.83	74.35	71.56	**99.11**	**98.71**	**99.11**	**98.36**	85.57	72.55	81.47	66.96	93.92	93.89	93.92	87.77
MH-Att	768+27	83.59	77.26	83.54	63.20	98.51	97.99	98.52	97.25	78.86	78.82	84.18	57.03	93.01	93.02	92.99	86.05
	768+768	85.77	74.38	88.52	66.27	98.21	97.57	98.23	96.73	82.66	73.53	83.78	61.18	93.92	93.89	93.92	87.77
BiGRU+Att	768+27	81.80	80.47	84.15	71.01	**98.81**	**98.43**	**98.82**	**97.82**	83.55	70.98	80.93	58.89	91.89	91.86	91.90	83.72
	768+768	82.19	71.56	81.78	66.35	98.51	97.86	98.52	97.24	79.07	70.42	82.55	62.62	93.72	93.70	93.72	87.42

Acc = Accuracy *F1-M* = F1 Macro *F1-W* = F1 Weighted *MCC* = Matthews Correlation Coefficient (%) **bold** = best per section & concatenation

Concatenation: 786+27 = 27 emotion dims ⊕ FinBERT 768 dims 768+768 = RoBERTa 768 dims ⊕ FinBERT 768 dims

Model abbreviations: SVM = Support Vector Machine RF = Random Forest LR = Logistic Regression XGB = XGBoost CatB = CatBoost LGBM = LightGBM BiGRU+Att = BiGRU + Attention Transf = Transformer Self-Att = Self-Attention Hier-Att = Hierarchical Attention MH-Att = Multi-Head Attention

Table 3.9 presents the performance of 16 machine learning, deep learning, and attention-based models using two concatenation configurations (768+27 and 768+768) across four financial sentiment datasets: PhraseBank 50%, PhraseBank 100%, Twitter, and Gold.

A. **result of 768+27** In the Machine Learning category, SVM, Random Forest, and XGBoost recorded the highest results on PhraseBank 100% (F1-W: 99.10%, MCC: 98.35%). On

PhraseBank 50%, XGBoost obtained the best scores (F1-W: 90.28%, MCC: 82.49%). SVM ranked first on Gold (F1-W: 93.12%, MCC: 86.21%) and Twitter (F1-W: 82.41%, MCC: 65.72%), making it the strongest ML model for the 768+27 representation.

In the Deep Learning category, GRU recorded the highest results on PhraseBank 100% (F1-W: 99.11%, MCC: 98.36%), while BiGRU ranked first on PhraseBank 50% (F1-W: 85.03%, MCC: 73.02%). LSTM led on Gold (F1-W: 92.91%, MCC: 85.84%). For Twitter, BiGRU obtained the highest results (F1-W: 77.66 %, MCC: 64.33%).

In the Attention category, BiGRU+Attention ranked first on PhraseBank 100% (F1-W: 98.82%, MCC: 97.82%). On PhraseBank 50%, BiGRU+Attention obtained the highest F1-W (84.15%), while Hierarchical Attention recorded the highest MCC (77.06%). MH-Attention led on Gold (F1-W: 92.99%, MCC: 86.05%). For Twitter, MH-Attention achieved the highest F1-W (84.18%), whereas Self-Attention obtained the highest MCC (66.40%).

B. **result of 768+768** In the Machine Learning category, CatBoost recorded the highest results on PhraseBank 100% (F1-W: 99.11%, MCC: 98.35%). SVM ranked first on PhraseBank 50% (F1-W: 90.40%, MCC: 82.71%), Gold (F1-W: 94.93%, MCC: 89.82%), and Twitter (F1-W: 84.92%, MCC: 70.62%), making it the strongest ML model for the 768+768 representation.

In the Deep Learning category, CNN-1D led on PhraseBank 100% (F1-W: 98.81%, MCC: 97.80%). LSTM recorded the best results on PhraseBank 50% (F1-W: 80.54%, MCC: 65.52%) and Twitter (F1-W: 84.93%, MCC: 65.97%), while BiGRU ranked first on Gold (F1-W: 94.33%, MCC: 88.61%).

In the Attention category, Hierarchical Attention ranked first on PhraseBank 100% (F1-W: 99.11%, MCC: 98.36%). Self-Attention led on PhraseBank 50% (F1-W: 85.77%, MCC: 68.71%) and Gold (F1-W: 95.34%, MCC: 90.63%), whereas MH-Attention recorded the best results on Twitter (F1-W: 83.78%, MCC: 61.18%).

Overall, PhraseBank 100% produced the highest scores, followed by Gold and PhraseBank 50%, while Twitter remained the most challenging dataset. Across the three model families, SVM showed the strongest ML results, BiGRU the strongest DL results, and Self-Attention the strongest attention-based results. The 768+768 representation generally outperformed the 768+27 representation, especially on the Gold and Twitter datasets.

Figure 3.4 presents the F1-Weighted and MCC scores of the best model per category (ML,DL ,ATT) on 768+768 representation.

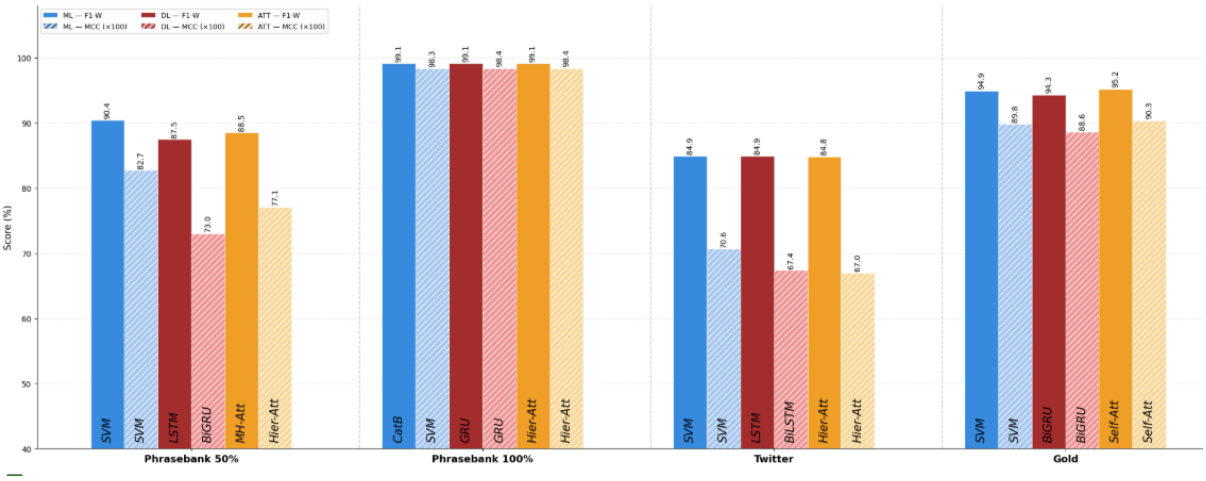


Figure 3.4. Best-performing model on 768+768 representation per category (ML, DL, ATT) — F1-Weighted and MCC scores across the four datasets (Concatenation Feature Extraction with Roberta-GoEmotion (27-768))

3.6.2 Concatenation Fine Tuning FinBERT with Roberta - GoEmotion (27-768)

This part presents the results of Approche 3 concatenation fine tuning FinBERT for financial sentiment classification using concatenated feature representations across four datasets. Two fusion strategies, namely 768+27 and 768+768, are evaluated using ML, DL, and ATT models. The corresponding results are summarized in Table 3.10.

Table 3.10. Approach 3 — Concatenation Fine tuning FinBERT: classification results across four datasets

Model	Concat.	Phrasebank 50%				Phrasebank 100%				Twitter				Gold			
		Acc	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*
Machine Learning (ML)																	
SVM	768+27	92.46	90.80	92.42	86.25	98.66	98.31	98.66	97.53	90.69	87.03	90.57	81.71	96.23	96.20	96.23	92.43
	768+768	92.66	90.98	92.64	86.66	96.43	95.38	96.42	93.45	**90.82**	**87.36**	**90.72**	**81.98**	**96.47**	**96.45**	**96.47**	**92.92**
RF	768+27	**93.08**	**91.27**	**93.03**	**87.38**	98.21	97.47	98.21	96.71	90.82	87.23	90.70	81.95	96.11	96.07	96.10	92.22
	768+768	**92.87**	**91.09**	**92.82**	**87.00**	95.98	94.89	95.93	92.60	90.73	87.09	90.60	81.76	96.11	96.07	96.10	92.22
LR	768+27	92.05	90.54	92.03	85.54	**99.11**	**98.74**	**99.11**	**98.36**	88.13	83.80	88.08	76.93	95.62	95.59	95.62	91.19
	768+768	92.45	91.10	92.45	86.36	95.98	94.94	95.96	92.61	88.64	84.62	88.60	77.94	95.38	95.35	95.62	91.19
XGB	768+27	92.87	91.16	92.84	87.02	**99.11**	98.73	99.10	98.35	**90.90**	**87.20**	**90.78**	**82.11**	96.11	96.08	96.10	92.20
	768+768	**92.87**	**91.02**	**92.83**	**87.01**	96.88	96.15	96.83	94.24	90.61	86.95	90.48	81.54	96.23	96.20	96.22	92.44
CatB	768+27	92.56	90.76	92.52	86.46	98.66	98.31	98.66	97.53	90.52	86.70	90.39	81.37	**96.59**	**96.57**	**96.59**	**93.17**
	768+768	92.35	90.52	92.31	86.09	96.43	95.34	96.40	93.43	90.69	87.03	90.56	81.70	96.35	96.32	96.34	92.70
LGBM	768+27	92.77	91.13	92.73	86.82	**99.11**	**98.73**	**99.10**	**98.35**	90.65	86.98	90.54	81.64	96.23	96.20	96.22	92.45
	768+768	92.45	90.79	92.43	86.27	**97.32**	**96.59**	**97.29**	**95.06**	90.78	87.06	90.65	81.86	**96.35**	**96.32**	**96.35**	**92.69**
Deep Learning (DL)																	
LSTM	768+27	**84.06**	**71.89**	**83.87**	**65.34**	97.32	96.23	97.33	95.09	**83.15**	**76.00**	**80.28**	**57.11**	96.35	96.35	96.34	92.70
	768+768	**82.31**	**76.32**	**78.16**	**58.27**	95.09	94.14	95.09	91.01	80.02	70.75	82.19	57.31	96.23	96.20	96.22	92.46
BiLSTM	768+27	82.17	71.69	81.95	59.97	97.77	96.67	97.79	95.93	83.76	70.15	79.90	56.76	**96.47**	**96.45**	**96.47**	**92.91**
	768+768	79.83	73.42	84.31	62.64	**96.43**	**95.34**	**96.40**	**93.43**	**85.50**	**72.71**	**78.40**	**60.78**	**96.59**	**96.57**	**96.59**	**93.15**
GRU	768+27	85.23	70.57	83.75	64.68	98.21	97.08	98.23	96.74	79.50	72.10	83.81	55.42	96.23	96.20	96.22	92.47

Continued on next page...

Table 3.10 — *(continued)*

Model	Concat.	Phrasebank 50%				Phrasebank 100%				Twitter				Gold			
		Acc	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*
	768+768	81.68	71.56	81.14	64.56	94.64	93.80	94.68	90.33	84.63	73.08	84.73	59.78	96.35	96.32	96.35	92.69
BiGRU	768+27	83.94	79.86	82.37	59.90	97.77	96.67	97.79	95.93	81.09	75.78	79.91	58.91	96.23	96.21	96.22	92.47
	768+768	**85.53**	**70.55**	**79.73**	**64.84**	95.54	94.63	95.58	91.96	84.35	74.48	78.36	62.12	96.11	96.08	96.11	92.17
CNN-1D	768+27	83.31	77.97	79.72	69.67	98.66	98.31	98.66	97.53	79.78	73.50	84.88	57.54	96.11	96.08	96.10	92.20
	768+768	79.20	70.50	79.79	60.47	96.88	95.80	96.88	94.34	83.00	70.64	83.38	59.81	96.47	96.44	96.47	92.94
Transf	768+27	87.58	78.44	86.93	63.55	**98.66**	**98.31**	**98.66**	**97.53**	81.56	70.85	79.29	61.60	**96.47**	**96.45**	**96.47**	**92.94**
	768+768	79.20	70.77	81.31	57.30	95.09	94.19	95.12	91.09	84.82	75.87	81.46	57.82	96.23	96.20	96.22	92.47
Attention (ATT)																	
Self-Att	768+27	**90.13**	**79.04**	**84.29**	**72.24**	**98.66**	**98.31**	**98.66**	**97.53**	83.72	83.29	82.59	65.24	**96.47**	**96.47**	**96.45**	**92.90**
	768+768	**80.63**	**71.29**	**83.43**	**60.39**	**98.66**	**97.90**	**98.67**	**97.54**	**86.94**	**77.80**	**81.07**	**62.13**	96.11	96.08	96.11	92.17
Hier-Att	768+27	87.20	84.01	81.66	77.14	96.88	95.82	96.88	94.29	87.22	73.92	84.08	64.04	96.23	96.20	96.23	92.42
	768+768	87.70	74.60	79.86	57.70	**98.66**	**97.90**	**98.67**	**97.54**	82.33	74.97	84.63	63.76	**96.23**	**96.20**	**96.23**	**92.42**
MH-Att	768+27	84.51	78.74	81.42	66.25	95.98	95.50	96.06	92.90	**84.51**	**74.20**	**85.92**	**74.82**	**96.47**	**96.46**	**96.44**	**92.98**
	768+768	84.35	70.63	80.69	63.37	97.77	97.51	97.79	95.97	83.93	74.45	83.37	64.80	95.01	95.00	95.02	90.15
BiGRU+Att	768+27	83.91	72.39	84.03	71.13	97.32	96.23	97.33	95.09	79.31	70.94	77.91	60.53	96.23	96.21	96.22	92.47
	768+768	79.72	71.81	80.31	61.93	95.98	94.94	95.96	92.61	80.34	76.19	79.48	60.83	96.11	96.08	96.10	92.19

Acc = Accuracy *F1-M* = F1 Macro *F1-W* = F1 Weighted *MCC* = Matthews Correlation Coefficient (%) **bold** = best per section & concatenation

Concatenation: 786+27 = 27 emotion dims ⊕ FinBERT 768 dims (fine-tuned) 768+768 = RoBERTa 768 dims ⊕ FinBERT 768 dims (fine-tuned)

Model abbreviations: SVM = Support Vector Machine RF = Random Forest LR = Logistic Regression XGB = XGBoost CatB = CatBoost LGBM = LightGBM BiGRU+Att = BiGRU + Attention Transf = Transformer Self-Att = Self-Attention Hier-Att = Hierarchical Attention MH-Att = Multi-Head Attention

Table 3.10 presents the classification performance of 16 machine learning, deep learning, and attention-based models using two fine-tuned FinBERT concatenation configurations (768+27 and 768+768) across four financial sentiment datasets: PhraseBank 100%, PhraseBank 50%, Twitter, and Gold.

A. **result of 768+27**

In the Machine Learning category, Logistic Regression, XGBoost, and LightGBM achieved the highest results on PhraseBank 100% (F1-W: 99.11%, MCC: 98.36%), Random Forest ranked first on PhraseBank 50% (F1-W: 93.03%, MCC: 87.38%), XGBoost led on Twitter (F1-W: 90.78%, MCC: 82.11%), and CatBoost achieved the best results on Gold (F1-W: 96.59%, MCC: 93.17%).

In the Deep Learning category, Transformer obtained the highest scores on PhraseBank 100% (F1-W: 98.66 %, MCC: 97.53 %),and is the best on PhraseBank 50% (F1-W: 86.93%, MCC: 63.55%), CNN-1D ranked first on Twitter (F1-W: 84.88%, MCC: 57.54%), while BiLSTM and Transformer achieved the highest results on Gold (F1-W: 96.47%, MCC: 92.94%).

In the Attention category, Self-Attention achieved the highest results on PhraseBank 100% (F1-W: 98.66%, MCC: 97.53%), Self-Attention on PhraseBank 50%(F1-W: 84.29%, MCC: 72.24%), while MH-Attention obtained the best scores on Twitter (F1-W: 85.92%,MCC:74.82%),Self-Att and MH-Att achieved the highest results on Gold (F1-W: 96.45%, MCC: 92.98%).

B. **Results of 768+768**

In the Machine Learning category, LightGBM achieved the highest results on PhraseBank 100% (F1-W: 97.29%, MCC: 95.06%), Random Forest and XGB ranked first on PhraseBank 50% (F1-W: 92.82%, MCC: 87.01%), SVM led on Twitter (F1-W: 90.72%, MCC: 81.98%), and CatBoost achieved the best results on Gold (F1-W: 96.34%, MCC: 92.70%).

In the Deep Learning category, CNN-1D achieved the highest scores on PhraseBank 100% (F1-W: 96.88%,MCC:94.34%), BiLSTM performed best on PhraseBank 50% (F1W:84.31%,MCC::62.64%) and Gold (F1-W: 96.59%, MCC: 93.15%), while GRU obtained the highest Twitter performance (F1-W: 84.73%, MCC: 59.78%).

In the Attention category,Self Attention and Hierarchical Attention achieved the highest results on PhraseBank 100% (F1-W: 98.67%, MCC: 97.54%) and Twitter (F1-W: 84.63%,MCC: 63.76), while self Attention performed best on PhraseBank 50% (F1-W: 83.43%,MCC: 60.39) ,and Hierarchical Attention achieved the highest result on Gold (F1-W: 96.23%,MCC: 92.42%).

Overall, PhraseBank 100% produced the highest classification scores, followed by Gold and PhraseBank 50%, while Twitter remained the most challenging dataset. The 768+27 configuration generally achieved better results than 768+768, suggesting that combining emotion-based features with fine-tuned FinBERT embeddings provides complementary information for financial sentiment classification.

Figure 3.5 presents the F1-Weighted and MCC scores of the best model per category (ML,DL ,ATT) on 768+768 representation.

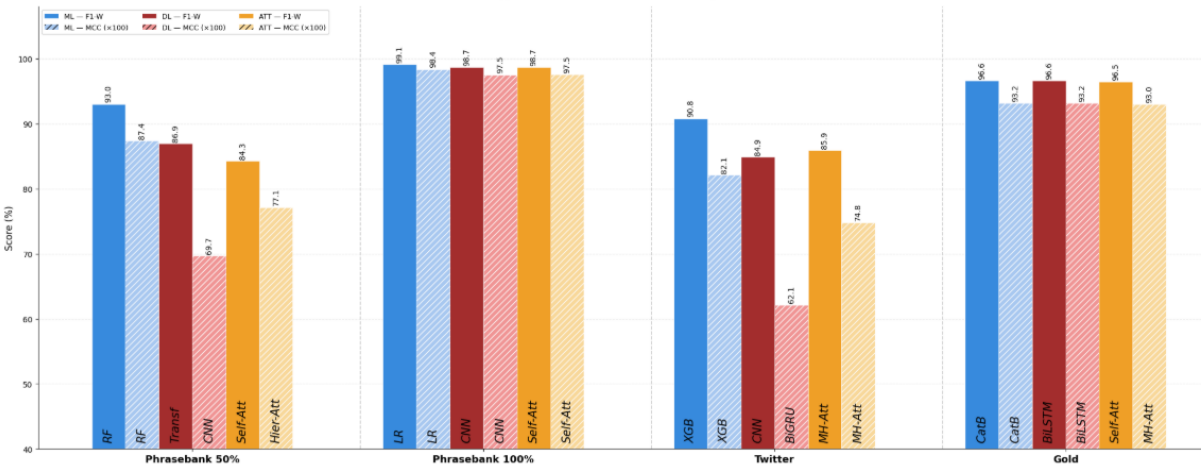


Figure 3.5. Best-performing model on 768+768 representation per category (ML, DL, ATT) — F1-Weighted and MCC scores across the four datasets (concatenation fine tuning FinBERT with Roberta-GoEmotion (27-768))

3.6.3 Concatenation Fine Tuning RoBERTa with Roberta - GoEmotion (27-768)

This part presents the results of Approach 3 concatenation Fine-tuning RoBERTa for financial sentiment classification using concatenated feature representations across four datasets. Two fusion strategies, namely 768+27 and 768+768, are evaluated using ML, DL, and ATT models. The corresponding results are summarized in Table 3.11.

Table 3.11. Approach 3 — Concatenation Fine tuning RoBERTa: classification results across four datasets

Model	Concat.	Phrasebank 50%				Phrasebank 100%				Twitter				Gold			
		Acc	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*
Machine Learning (ML)																	
SVM	768+27	92.46	91.48	92.46	86.33	98.66	98.30	98.65	97.53	92.75	90.48	92.73	85.94	95.86	95.84	95.84	91.69
	768+768	92.36	91.64	92.35	86.11	99.11	98.73	99.10	98.35	92.41	90.10	92.40	85.30	96.23	96.20	96.23	92.43
RF	768+27	92.56	91.44	92.54	86.47	98.66	98.30	98.65	97.53	92.87	90.64	92.84	86.12	96.35	96.33	96.35	92.69
	768+768	92.46	91.44	92.44	86.26	98.55	98.30	98.65	97.53	**92.83**	**90.57**	**92.80**	**86.03**	96.11	96.08	96.10	92.20
LR	768+27	90.19	88.91	90.18	82.22	99.11	99.15	99.10	98.35	90.69	87.88	90.69	82.05	95.13	95.10	95.13	90.22
	768+768	90.50	89.50	90.48	82.71	98.66	98.30	98.65	97.53	91.28	88.73	91.28	83.13	94.65	94.62	94.65	89.23
XGB	768+27	92.46	91.33	92.43	86.24	99.11	99.15	99.10	98.35	92.62	90.30	92.57	85.58	**96.11**	**96.08**	**96.10**	**92.19**
	768+768	92.25	91.06	92.21	85.85	**99.11**	**99.15**	**99.10**	**98.35**	**92.83**	**90.52**	**92.79**	**86.02**	**96.23**	**96.20**	**96.22**	**92.44**
CatB	768+27	**92.87**	**91.82**	**92.85**	**87.03**	98.55	98.30	98.65	97.53	92.66	90.33	92.62	85.69	96.35	96.32	96.35	92.68
	768+768	**92.67**	**91.56**	**92.65**	**86.65**	98.66	98.30	98.65	97.53	92.75	90.46	92.71	85.88	96.35	96.32	96.35	92.69
LGBM	768+27	92.46	91.45	92.44	86.27	**99.55**	**99.58**	**99.55**	**99.17**	**92.96**	**90.75**	**92.92**	**86.26**	**96.47**	**96.45**	**96.47**	**92.94**
	768+768	92.36	91.42	92.33	86.06	**99.11**	98.73	99.10	98.35	92.75	90.46	92.71	85.85	**96.35**	**96.32**	**96.35**	**92.68**
Deep Learning (DL)																	
LSTM	768+27	84.72	70.89	81.04	68.33	98.66	98.30	98.65	97.53	79.27	68.97	78.30	66.44	96.23	96.21	96.23	92.42
	768+768	81.67	76.11	80.90	61.04	98.66	98.30	98.66	97.53	**86.14**	**69.78**	**84.62**	**58.99**	95.86	95.84	95.86	91.69

Continued on next page...

Table 3.11 — *(continued)*

Model	Concat.	Phrasebank 50%				Phrasebank 100%				Twitter				Gold			
		Acc	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*
BiLSTM	768+27	84.04	76.85	79.51	60.38	98.66	98.30	98.65	97.53	83.76	70.15	79.90	56.76	95.99	95.95	95.98	91.98
	768+768	81.24	71.29	81.14	57.97	99.11	98.73	99.10	98.35	81.54	74.44	80.57	66.84	96.11	96.08	96.10	92.18
GRU	768+27	80.06	79.08	86.75	68.73	98.21	97.90	98.22	96.71	78.22	70.44	78.39	55.02	**96.23**	**96.20**	**96.23**	**92.42**
	768+768	79.65	73.33	82.08	64.62	98.21	97.90	98.22	96.71	82.16	71.27	82.78	57.49	95.86	95.83	95.86	91.75
BiGRU	768+27	86.10	70.84	79.38	60.26	98.21	97.90	98.22	96.71	84.72	68.20	82.34	62.97	**96.35**	**96.33**	**96.35**	**92.66**
	768+768	85.29	74.13	79.35	60.75	**99.11**	**98.73**	**99.10**	**98.35**	80.23	76.35	80.98	60.21	**96.11**	**96.08**	**96.10**	**92.22**
CNN-1D	768+27	80.64	72.48	83.86	67.14	98.66	98.30	98.66	97.53	81.20	68.97	80.77	56.13	96.23	96.21	96.23	92.42
	768+768	80.30	70.55	79.31	62.24	99.11	98.73	99.10	98.35	84.83	74.03	78.72	64.57	95.74	95.70	95.73	91.51
Transf	768+27	83.54	72.44	79.63	70.34	98.66	98.30	98.65	97.53	78.88	68.90	80.58	63.47	96.35	96.32	96.35	92.68
	768+768	81.81	70.48	83.22	57.08	**99.11**	**98.73**	**99.10**	**98.35**	81.89	69.27	78.53	60.54	95.13	95.08	95.12	90.35
Attention (ATT)																	
Self-Att	768+27	85.05	78.69	85.59	68.40	**99.11**	**98.73**	**99.10**	**98.35**	88.32	79.16	82.05	77.73	**96.72**	**96.69**	**96.71**	**99.32**
	768+768	**86.69**	**75.41**	**83.77**	**58.72**	**99.11**	**98.73**	**99.10**	**98.35**	**85.53**	**75.05**	**81.70**	**66.80**	95.99	95.96	95.99	91.93
Hier-Att	768+27	**87.49**	**74.77**	**88.63**	**67.63**	98.66	98.30	98.65	97.53	**89.65**	**78.68**	**83.44**	**72.24**	**96.59**	**96.57**	**96.59**	**99.29**
	768+768	85.26	70.57	80.48	65.56	97.77	97.25	97.78	95.92	80.75	72.59	82.93	61.21	**96.23**	**96.20**	**96.23**	**92.42**
MH-Att	768+27	83.77	80.81	86.01	75.64	98.21	97.90	98.22	96.71	88.95	82.55	82.20	63.53	**96.23**	**96.20**	**96.22**	**99.24**
	768+768	84.82	77.23	81.07	67.26	**99.11**	**98.73**	**99.10**	**98.35**	80.92	74.00	83.79	59.69	94.89	94.88	94.90	89.87
BiGRU+Att	768+27	83.04	76.61	84.81	61.38	98.66	98.30	98.65	97.53	83.33	71.13	82.97	55.47	**96.35**	**96.33**	**96.35**	**92.66**
	768+768	**83.23**	**74.73**	**81.16**	**61.76**	**99.11**	**98.73**	**99.10**	**98.35**	83.56	70.10	78.15	58.55	**96.35**	**96.33**	**96.35**	**92.67**

Acc = Accuracy *F1-M* = F1 Macro *F1-W* = F1 Weighted *MCC* = Matthews Correlation Coefficient (%) **bold** = best per section & concatenation

Concatenation: 786+27 = 27 emotion dims ⊕ RoBERTa 768 dims (fine-tuned) 768+768 = RoBERTa 768 dims ⊕ RoBERTa 768 dims (fine-tuned)

Model abbreviations: SVM = Support Vector Machine RF = Random Forest LR = Logistic Regression XGB = XGBoost CatB = CatBoost LGBM = LightGBM BiGRU+Att = BiGRU + Attention Transf = Transformer Self-Att = Self-Attention Hier-Att = Hierarchical Attention MH-Att = Multi-Head Attention

Table 3.11 presents the classification performance of 16 machine learning, deep learning,and attention-based models using two fine-tuned RoBERTa concatenation configurations(768+27 and 768+768) across four financial sentiment datasets: PhraseBank 100%, PhraseBank 50%, Twitter, and Gold.

A. **result of 768+27**

In the Machine Learning category, LightGBM achieved the best results on PhraseBank 100% (F1-W: 99.55%, MCC: 99.17%), CatBoost ranked first on PhraseBank 50% (F1-W: 92.85%, MCC: 87.03%), while LightGBM also led on Twitter (F1-W: 92.92%, MCC: 86.26%) and Gold (F1-W: 96.47%, MCC: 92.94%).

In the Deep Learning category, GRU achieved the best performance on PhraseBank 50% (F1-W: 86.75%, MCC: 68.73%), BiGRU ranked first on Twitter (F1-W: 82.34%, MCC:62.97%), while several models achieved similar top results on PhraseBank 100% (F1-W: 98.65%,MCC:97.53%) and Gold (F1-W = 96.35%, MCC = 92.68%).

In the Attention category, Self-Attention achieved the highest results on PhraseBank 100% (F1-W: 99.10%, MCC: 98.35%), Hierarchical Attention performed best on PhraseBank 50% (F1-W: 88.63%,MCC:67.63%), while Hierarchical Attention and Self Attention achieved the strongest results on Twitter (F1-W: 83.44%, MCC: 77.73%) and

Gold (F1-W: 96.71%, MCC: 99.32%).

B. **result of 768+768**

In the Machine Learning category, CatBoost achieved the best performance on PhraseBank 50% (F1-W: 92.65%, MCC: 86.65%), SVM, XGBoost, and LightGBM led on PhraseBank 100% (F1-W: 99.10%, MCC: 98.35%), XGBoost ranked first on Twitter (F1-W: 92.79%, MCC: 86.02%), while CatBoost and LightGBM achieved the best results on Gold (F1-W: 96.35%, MCC: 92.69%).

In the Deep Learning category, Transformer achieved the highest F1-W on PhraseBank 50% (83.22%, MCC: 57.08%), GRU obtained the best Twitter performance (F1-W: 82.78%, MCC: 57.49%), and several models achieved identical top results on PhraseBank 100% (F1-W: 99.10%, MCC: 98.35%). BiGRU ,BiLSTM achieved the best results on Gold (F1-W: 96.35%, MCC: 92.68%).

In the Attention category, Self-Attention and MH-Attention achieved the highest results on PhraseBank 100% (F1-W: 99.10%, MCC: 98.35%). Self-Attention performed best on PhraseBank 50% (F1-W: 83.77%,MCC:58.72 %) and BiGRU+Att on Gold (F1-W: 96.35%, MCC: 91.93%), while MH-Attention achieved the highest Twitter performance (F1-W: 83.79%, MCC: 59.69%).

Figure 3.6 presents the F1-Weighted and MCC scores of the best model per category (ML,DL ,ATT) on 768+768 representation.

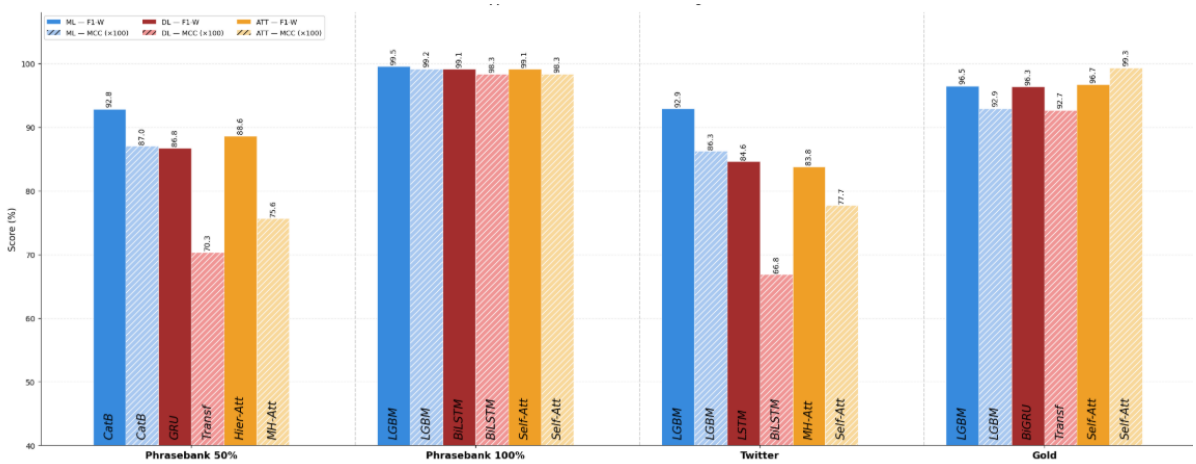


Figure 3.6. Best-performing model on 768+768 representation per category (ML, DL, ATT) — F1-Weighted and MCC scores across the four datasets (concatenation fine tuning RoBERTa with Roberta - GoEmotion (27-768))

3.6.4 Concatenation Fine Tuning DeBERTa with Roberta - GoEmotion (27-768)

This part presents the results of Approach 3 concatenation Fine-tuning DeBERTa for financial sentiment classification using concatenated feature representations across four datasets. Two fusion strategies, namely 768+27 and 768+768, are evaluated using ML, DL, and ATT models. The corresponding results are summarized in Table 3.12.

Table 3.12. Approach 3 — Concatenation Fine tuning DeBERTa: classification results across four datasets

Model	Concat.	Phrasebank 50%				Phrasebank 100%				Twitter				Gold			
		Acc	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*
Machine Learning (ML)																	
SVM	768+27	89.46	87.37	89.46	80.90	98.66	98.65	98.66	97.53	87.88	83.80	87.76	76.16	95.99	95.96	95.98	91.95
	768+768	89.26	87.23	89.27	80.63	**99.11**	**98.73**	**99.10**	**98.35**	88.05	84.18	87.96	76.58	95.99	95.96	95.98	91.94
RF	768+27	89.57	87.68	89.59	81.23	**99.11**	**98.73**	**99.10**	**98.35**	**88.09**	**84.20**	**88.03**	**76.73**	95.38	95.36	95.37	90.76
	768+768	**89.67**	**87.67**	**89.68**	**81.39**	**99.11**	**98.73**	**99.10**	**98.35**	88.01	84.19	87.96	76.62	95.62	95.60	95.61	91.27
LR	768+27	86.36	85.00	86.41	75.51	98.66	98.65	98.66	97.53	84.28	79.68	84.43	70.30	93.80	93.78	93.79	87.52
	768+768	87.40	86.12	87.47	77.60	**99.11**	**98.73**	**99.10**	**98.35**	84.99	80.36	85.00	71.04	94.04	94.00	94.04	88.01
XGB	768+27	89.46	87.33	89.45	80.87	**99.11**	**98.73**	**99.10**	**98.35**	87.80	83.77	87.68	75.98	95.86	95.84	95.86	91.72
	768+768	89.46	87.59	89.46	80.93	**99.11**	**98.73**	**99.10**	**98.35**	88.22	84.35	88.13	76.89	95.99	95.95	95.49	90.99
CatB	768+27	**89.67**	**87.68**	**89.68**	**81.35**	**99.11**	**98.73**	**99.10**	**98.35**	88.09	84.14	87.99	76.64	95.74	95.71	95.73	91.48
	768+768	89.26	87.19	89.26	80.59	**99.11**	**98.73**	**99.10**	**98.35**	88.26	84.36	88.16	76.97	95.74	95.71	95.73	91.48
LGBM	768+27	89.36	87.39	89.36	80.76	**99.11**	**98.73**	**99.10**	**98.35**	88.01	83.94	87.89	76.41	**95.86**	**95.84**	**95.86**	**91.72**
	768+768	89.36	87.14	89.38	80.81	**99.11**	**98.73**	**99.10**	**98.35**	**88.39**	**84.52**	**88.29**	**77.18**	**95.62**	**95.58**	**95.61**	**91.24**
Deep Learning (DL)																	
LSTM	768+27	82.09	79.09	83.36	58.72	**98.66**	**98.31**	**98.66**	**97.53**	86.65	75.99	80.48	55.74	95.74	95.71	95.74	91.47
	768+768	84.10	76.18	82.74	57.99	**98.66**	**98.31**	**98.66**	**97.53**	86.98	76.03	79.17	58.50	**95.99**	**95.95**	**95.98**	**91.98**
BiLSTM	768+27	82.60	72.43	83.06	67.05	**98.66**	**98.31**	**98.66**	**97.53**	**82.22**	**71.43**	**82.52**	63.34	95.13	95.11	95.12	90.26
	768+768	79.57	70.19	84.31	61.43	**98.66**	**98.31**	**98.66**	**97.53**	79.60	76.69	83.40	57.45	95.62	95.58	95.61	91.29
GRU	768+27	82.58	71.31	83.69	61.98	**98.66**	**98.31**	**98.66**	**97.53**	**83.69**	**72.79**	**81.38**	**61.89**	95.74	95.70	95.74	91.47
	768+768	79.68	69.28	79.00	61.73	**98.66**	**98.31**	**98.66**	**97.53**	86.56	70.65	81.18	66.29	95.74	95.70	95.24	90.57
BiGRU	768+27	83.83	71.55	85.04	58.06	**98.66**	**98.31**	**98.66**	**97.53**	79.77	74.09	81.28	60.31	95.38	95.33	95.36	90.82
	768+768	85.77	76.15	81.90	67.13	**98.66**	**98.31**	**98.66**	**97.53**	84.72	72.40	79.57	66.65	95.26	95.24	95.98	91.94
CNN-1D	768+27	84.65	77.00	84.25	66.40	98.21	97.90	98.22	96.71	84.66	69.43	79.19	60.13	95.99	95.96	95.98	91.93
	768+768	86.82	71.32	80.23	66.71	98.66	98.31	98.66	97.53	83.31	71.95	81.84	57.99	94.40	94.35	94.39	88.83
Transf	768+27	81.96	70.74	81.67	62.71	**98.66**	**98.31**	**98.65**	**97.53**	79.54	69.57	81.40	55.10	**95.99**	**95.95**	**95.98**	**92.00**
	768+768	79.74	72.53	83.39	60.96	**98.66**	**98.31**	**98.66**	**97.53**	83.71	72.18	78.68	60.53	95.38	95.33	95.36	90.82
Attention (ATT)																	
Self-Att	768+27	89.46	76.65	87.47	79.79	**99.11**	**98.73**	**99.10**	**98.35**	87.52	73.81	87.97	73.47	**96.35**	**96.33**	**96.35**	**92.67**
	768+768	87.35	72.69	83.69	59.45	**99.11**	**98.73**	**99.10**	**98.35**	81.25	73.75	81.62	60.97	95.86	95.83	95.86	99.38
Hier-Att	768+27	90.65	82.87	81.81	66.66	**99.11**	**98.73**	**99.10**	**98.35**	85.18	75.68	86.17	70.07	95.99	95.96	95.98	91.94
	768+768	81.73	75.99	79.05	58.39	**99.11**	**98.73**	**99.10**	**98.35**	82.76	70.13	80.29	57.68	95.38	95.34	95.37	99.35
MH-Att	768+27	87.16	77.72	82.29	79.88	98.66	98.08	98.67	97.54	**87.25**	**75.34**	**87.64**	**74.17**	**96.23**	**96.20**	**96.23**	**92.43**
	768+768	**87.14**	**74.46**	**85.02**	**66.09**	**99.11**	**98.73**	**99.10**	**98.35**	**87.97**	**76.72**	**84.63**	**58.99**	**95.86**	**95.84**	**95.86**	**99.27**
BiGRU+Att	768+27	**81.83**	**74.97**	**86.69**	**67.27**	**98.66**	**98.31**	**98.66**	**97.53**	81.91	68.19	78.61	63.32	95.01	94.97	95.00	90.03
	768+768	79.48	75.53	83.07	62.71	**98.66**	**98.31**	**98.66**	**97.53**	84.08	73.31	84.21	65.88	95.13	95.08	95.12	90.31

Acc = Accuracy *F1-M* = F1 Macro *F1-W* = F1 Weighted *MCC* = Matthews Correlation Coefficient (%) **bold** = best per section & concatenation

Concatenation: 768+27 = 27 emotion dims ⊕ DeBERTa 768 dims (fine-tuned) 768+768 = RoBERTa 768 dims ⊕ DeBERTa 768 dims (fine-tuned)

Model abbreviations: SVM = Support Vector Machine RF = Random Forest LR = Logistic Regression XGB = XGBoost CatB = CatBoost LGBM = LightGBM BiGRU+Att = BiGRU + Attention Transf = Transformer Self-Att = Self-Attention Hier-Att = Hierarchical Attention MH-Att = Multi-Head Attention

Table 3.12 presents the classification performance of 16 machine learning, deep learning,and attention based models using two fine-tuned DeBERTa concatenation

configurations(768+27 and 768+768) across four financial sentiment datasets: PhraseBank 100%, PhraseBank 50%,Twitter, and Gold.

A. **result of 768+27**

In the Machine Learning category, CatBoost achieved the best results on PhraseBank 50% (F1-W: 89.68%, MCC: 81.35%), RF,XGB,CatB,LightGBM on PhraseBank 100% (F1-W: 99.10%, MCC: 98.35%), Random Forest on Twitter (F1-W: 88.03%, MCC: 76.73%), and SVM on Gold (F1-W: 95.98%, MCC: 91.95%).

In the Deep Learning category, BiGRU+Attention achieved the best performance on PhraseBank 50% (F1-W: 86.69%), BiGRU on Twitter (F1-W: 85.04 %, MCC: 58.06%), while several models obtained similar top results on PhraseBank 100% (F1-W: 98.66%, MCC: 97.53%). Transformer achieved the highest performance on Gold (F1-W: 95.98%, MCC: 92.00%).

In the Attention category,self Attention performed best on PhraseBank 50% (F1-W: 87.47%, MCC: 79.79%), Self Attention ,Hierarchical Attention on PhraseBank 100% (F1-W: 99.10%, MCC: 98.35%) and Gold (F1-W: 96.35%, MCC: 92.67%), while MH-Attention achieved the strongest Twitter results (F1-W: 87.64%, MCC: 74.17%).

B. **result of 768+768**

In the Machine Learning category, Random Forest achieved the best results on PhraseBank 50% (F1-W: 89.68%, MCC: 81.39%), all ML models obtained identical top performance on PhraseBank 100% (F1-W: 99.10%, MCC: 98.35%), LightGBM ranked first on Twitter (F1-W: 88.29%, MCC: 77.18%), and SVM achieved the highest Gold performance (F1-W: 95.98%, MCC: 91.94%).

In the Deep Learning category, BiLSTM achieved the best performance on PhraseBank 50% (F1-W: 84.31% , MCC:61.43%), BiLSTM obtained the strongest Twitter results (F1-W: 83.40%, MCC: 57.45%), LSTM achieved the highest Gold performance (F1-W: 95.98%, MCC: 91.98%), while several DL models reached similar top results on PhraseBank 100% (F1-W: 98.66%, MCC: 97.53%).

In the Attention category, all attention models achieved identical top performance on PhraseBank 100% (F1-W: 99.10%, MCC: 98.35%), MH-Attention obtained the best Twitter results (F1-W: 84.63%, MCC: 58.99%), and Self-Attention achieved the highest Gold performance (F1-W: 95.86%, MCC: 99.27%).

Overall, PhraseBank 100% yielded the highest classification performance, whereas Twitter remained the most challenging dataset.The 768+768 configuration showed a slight advantage on Twitter, while producing comparable results to 768+27 on PhraseBank 100% and Gold.

Figure 3.7 presents the F1-Weighted and MCC scores of the best model per category (ML,DL ,ATT) on 768+768 representation.

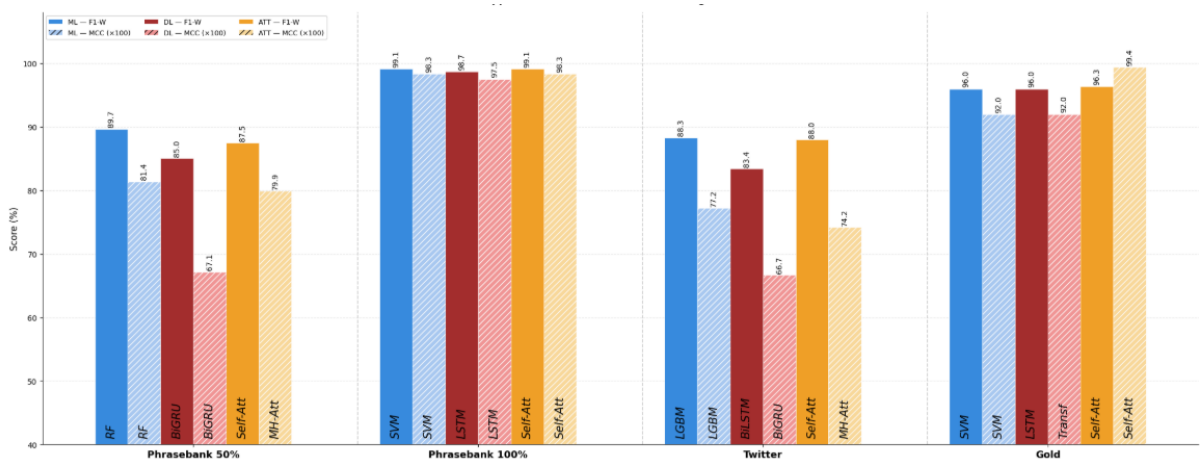


Figure 3.7. Best-performing model on 768+768 representation per category (ML, DL, ATT) — F1-Weighted and MCC scores across the four datasets (concatenation fine tuning DeBERTa with Roberta-GoEmotion (27-768))

3.6.5 Concatenation Fine Tuning XLM-RoBERTa with Roberta - GoEmotion (27-768)

This part presents the results of Approach 3 Concatenation Fine tuning XLM-RoBERTa for financial sentiment classification using concatenated feature representations across four datasets. Two fusion strategies, namely 768+27 and 768+768, are evaluated using ML, DL, and ATT models. The corresponding results are summarized in Table 3.13.

Table 3.13. Approach 3 — Concatenation Fine tuning XLM-RoBERTa: classification results across four datasets

Model	Concat.	Phrasebank 50%				Phrasebank 100%				Twitter				Gold			
		Acc	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*
Machine Learning (ML)																	
SVM	768+27	**91.01**	**89.96**	**91.06**	**84.05**	96.43	95.38	96.42	93.45	90.06	86.79	90.06	80.82	95.74	95.70	95.73	91.51
	768+768	91.43	90.34	91.45	84.63	96.88	96.22	96.87	94.25	90.40	87.14	90.36	81.35	95.86	95.83	95.85	91.76
RF	768+27	90.60	89.21	90.60	83.02	95.98	94.94	95.93	92.60	**91.07**	**87.95**	**90.97**	**82.48**	95.38	95.36	95.37	90.76
	768+768	90.60	89.21	90.61	83.07	95.98	94.94	95.93	92.60	90.78	87.68	90.69	81.94	95.74	95.70	95.73	91.51
LR	768+27	89.36	88.30	89.42	81.16	95.98	94.94	95.96	92.61	88.39	84.96	88.43	77.79	95.50	95.48	95.49	90.99
	768+768	89.26	88.40	89.30	80.82	95.98	95.37	95.97	92.61	88.34	84.79	88.36	77.58	95.62	95.58	94.88	89.75
XGB	768+27	90.81	89.70	90.81	83.37	96.88	96.15	96.83	94.24	90.86	87.67	90.74	82.03	95.74	95.70	95.73	91.49
	768+768	90.91	89.60	90.92	83.58	97.32	96.59	97.29	95.06	**90.94**	**87.89**	**90.83**	**82.20**	95.74	95.70	95.73	91.51
CatB	768+27	90.91	89.69	90.92	83.61	96.43	95.34	96.40	93.43	90.73	87.47	90.61	81.79	**95.86**	**95.83**	**95.86**	**91.73**
	768+768	91.22	89.87	91.23	84.15	96.43	95.78	96.41	93.41	90.90	87.72	90.79	82.13	**95.86**	**95.83**	**95.86**	**91.73**
LGBM	768+27	91.74	90.77	91.75	85.11	**97.32**	**96.59**	**97.29**	**95.06**	90.94	87.79	90.83	82.18	**95.62**	**95.60**	**95.61**	**91.24**
	768+768	91.53	90.52	91.53	84.71	**97.77**	**97.03**	**97.75**	**95.89**	**90.99**	**87.78**	**90.88**	**82.29**	**95.86**	**95.83**	**95.86**	**91.73**
Deep Learning (DL)																	
LSTM	768+27	85.59	71.18	80.42	68.08	95.98	94.94	95.56	92.61	86.63	72.26	78.58	58.49	95.99	95.95	95.98	91.98
	768+768	85.08	76.46	84.48	66.29	95.54	95.02	95.56	91.90	83.57	74.58	80.08	67.45	95.62	95.96	95.61	91.25
BiLSTM	768+27	81.89	77.52	82.19	66.05	94.64	93.80	94.68	90.33	80.37	70.50	82.37	59.32	95.74	95.71	95.73	91.51
	768+768	82.86	74.96	81.81	60.43	95.54	94.98	95.54	91.81	84.28	70.88	84.76	58.48	95.74	95.71	95.74	91.45
GRU	768+27	84.89	71.80	80.81	61.27	**95.54**	**95.02**	**95.56**	**91.90**	78.95	71.16	84.62	62.16	95.62	95.58	95.61	91.27

Continued on next page...

Table 3.13 — *(continued)*

Model	Concat.	Phrasebank 50%				Phrasebank 100%				Twitter				Gold			
		Acc	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*	*Acc*	*F1-M*	*F1-W*	*MCC*
	768+768	85.80	75.89	80.88	67.83	**96.43**	**95.78**	**96.41**	**93.41**	83.14	75.47	83.56	66.30	95.99	95.96	95.98	91.98
BiGRU	768+27	86.74	74.74	81.92	62.68	95.09	94.19	95.12	91.09	84.71	68.71	83.40	56.31	**95.74**	**95.70**	**95.73**	**91.51**
	768+768	85.21	69.13	84.53	58.40	95.54	94.98	95.54	91.81	84.73	72.89	81.23	57.36	**95.74**	**95.71**	**95.74**	**91.47**
CNN-1D	768+27	83.04	77.67	82.68	67.61	94.64	92.81	94.60	90.12	85.08	75.52	78.21	57.77	95.86	95.83	95.86	91.75
	768+768	83.42	75.67	79.91	61.59	94.64	93.40	94.17	89.24	80.82	71.68	84.88	59.66	95.86	95.83	95.86	91.73
Transf	768+27	80.33	77.74	86.67	61.18	**95.54**	**95.02**	**95.56**	**91.90**	85.63	74.69	83.46	63.06	95.50	95.45	95.48	91.08
	768+768	85.97	74.68	80.58	63.65	95.54	94.98	95.54	91.81	81.70	74.99	83.33	62.06	**95.99**	**95.96**	**95.98**	**91.94**
Attention (ATT)																	
Self-Att	768+27	88.89	84.46	82.08	67.11	**96.43**	**95.78**	**96.41**	**93.41**	**83.72**	**81.78**	**87.17**	**66.41**	96.11	96.07	96.10	92.24
	768+768	86.97	70.52	82.84	64.78	96.88	94.98	96.22	94.25	86.32	72.29	80.82	59.11	**95.86**	**95.83**	**95.86**	**91.75**
Hier-Att	768+27	**83.56**	**82.16**	**83.93**	**68.85**	95.09	94.19	95.12	91.09	84.79	78.69	84.94	64.70	95.99	95.95	95.98	92.02
	768+768	81.81	70.63	81.14	61.46	95.98	94.19	95.42	92.67	87.21	72.54	81.55	57.33	95.50	95.46	95.49	90.96
MH-Att	768+27	89.10	76.93	83.01	64.18	95.54	94.98	95.54	91.81	87.96	76.92	81.76	66.81	**96.23**	**96.20**	**96.23**	**92.42**
	768+768	**85.00**	**73.18**	**85.51**	**64.30**	**96.88**	**95.78**	**96.28**	**94.39**	**83.78**	**74.72**	**85.62**	**65.55**	**95.86**	**95.82**	**95.85**	**91.76**
BiGRU+Att	768+27	83.11	77.62	83.78	71.14	95.98	94.94	95.56	92.61	86.84	73.36	77.67	55.22	**95.74**	**95.70**	**95.73**	**91.51**
	768+768	86.82	73.35	80.82	66.48	95.54	95.02	95.56	91.90	80.10	76.84	80.24	66.00	**95.99**	**95.96**	**95.98**	**91.95**

Acc = Accuracy *F1-M* = F1 Macro *F1-W* = F1 Weighted *MCC* = Matthews Correlation Coefficient (%) **bold** = best per section & concatenation

Concatenation: 768+27 = 27 emotion dims ⊕ XLM-RoBERTa 768 dims (fine-tuned) 768+768 = RoBERTa 768 dims ⊕ XLM-RoBERTa 768 dims (fine-tuned)

Model abbreviations: SVM = Support Vector Machine RF = Random Forest LR = Logistic Regression XGB = XGBoost CatB = CatBoost LGBM = LightGBM BiGRU+Att = BiGRU + Attention Transf = Transformer Self-Att = Self-Attention Hier-Att = Hierarchical Attention MH-Att = Multi-Head Attention

Table 3.13 presents the classification performance of 16 machine learning, deep learning,and attention based models using two fine-tuned DeBERTa concatenation configurations(768+27 and 768+768) across four financial sentiment datasets: PhraseBank 100%, PhraseBank 50%,Twitter, and Gold.

A. **result of 786+27**

In the Machine Learning category, LightGBM achieved the best results on PhraseBank 50% (F1-W: 91.75%, MCC: 85.11%), PhraseBank 100% (F1-W: 97.29%, MCC: 95.06%), CatB on Gold (F1-W: 95.86%, MCC: 91.73%), while Random Forest ranked first on Twitter (F1-W: 90.97%, MCC: 82.48%).

In the Deep Learning category, Transformer achieved the best performance on PhraseBank 50% (F1-W: 86.67%), BiGRU+Attention obtained the highest MCC (71.14%), GRU ranked first on Twitter (F1-W: 84.62%,MCC:62.16), while Transformer achieved the highest Gold performance (F1-W: 95.48%, MCC: 91.08%). Several DL models obtained similar results on PhraseBank 100% (F1-W: 95.56%,MCC:92.61).

In the Attention category, Hierarchical Attention achieved the best results on PhraseBank 50% (F1-W: 83.93%, MCC: 68.85%), Self-Attention ranked first on PhraseBank 100% (F1-W: 96.41%, MCC: 93.41%) and Twitter (F1-W: 87.17%,MCC: 66.81%), while MH-Attention obtained the highest Gold performance (F1-W: 96.23%, MCC: 92.42%).

B. **result of 768+768**

In the Machine Learning category, LightGBM achieved the best results on PhraseBank

100% (F1-W: 97.75%, MCC: 95.89%), Twitter (F1-W: 90.88%, MCC: 82.29%), and Gold (F1-W: 95.86%, MCC: 91.76%), and PhraseBank 50% (F1-W: 91.53%, MCC: 84.71%).

In the Deep Learning category, BiGRU achieved the best performance on PhraseBank 50% (F1-W: 84.53%, MCC: 67.83%), CNN-1D ranked first on Twitter (F1-W: 84.88%,MCC: 62.06%), and Transf achieved the highest Gold performance (F1-W: 95.98%, MCC: 91.94%). and GRU on PhraseBank 100% (F1-W: 96.41%, MCC: 93.41%).

In the Attention category, MH-Attention achieved the best results on PhraseBank 50% (F1-W: 85.51%,MCC: 64.30%), PhraseBank 100% (F1-W: 96.28%, MCC: 94.39%), and Twitter (F1-W: 85.62%, MCC: 65.55%), while BiGRU+Attention obtained the highest Gold performance (F1-W: 95.98%, MCC: 91.95%).

Figure 3.8 presents the F1-Weighted and MCC scores of the best model per category (ML,DL ,ATT) on 768+768 representation.

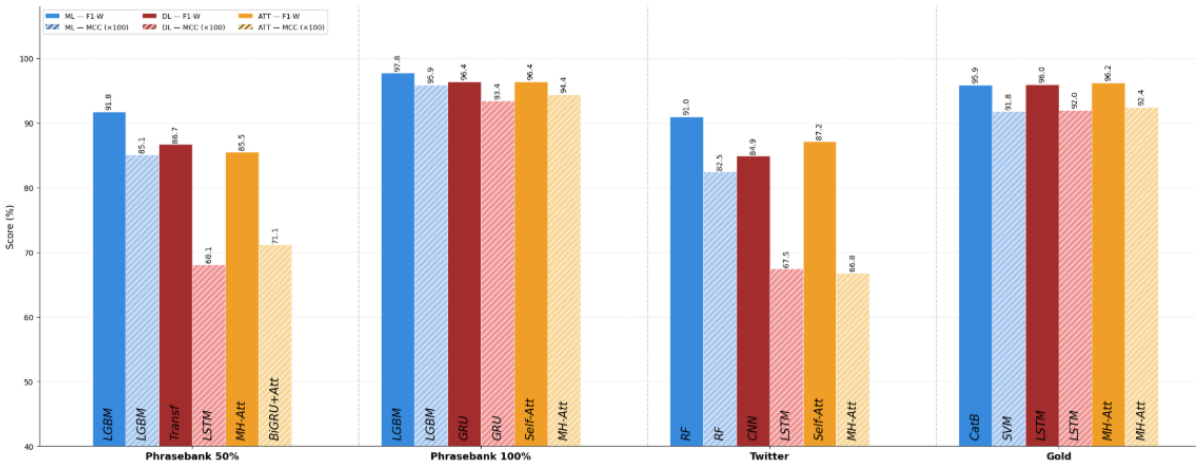


Figure 3.8. Best-performing model on 768+768 representation per category (ML, DL, ATT) — F1-Weighted and MCC scores across the four datasets (concatenation fine tuning XLM-RoBERTa with Roberta-GoEmotion (27-768))

3.6.6 Discussion Results of Approach 3

The results of Approach 3 indicate that the 768+768 representation generally outperformed the 768+27 representation across the four fine-tuned transformer models. Combining two high-dimensional contextual embeddings preserved richer semantic and financial information than integrating the 27-dimensional emotion vector, leading to more consistent improvements in F1-Weighted and MCC scores, particularly on Twitter and Gold.

- Characteristics remained a major factor affecting performance. PhraseBank 100% consistently achieved the highest scores due to complete annotator agreement, while Gold also produced strong and stable results because of its high-quality annotations. In contrast, Twitter remained the most challenging dataset because of its noisy and informal language, resulting in lower performance across all configurations.

- Among the four fine-tuned transformers, RoBERTa achieved the strongest overall performance, followed closely by FinBERT. Although FinBERT benefited from its

financial-domain pretraining, RoBERTa consistently obtained slightly higher results across most datasets and model families. DeBERTa and XLM-RoBERTa also produced competitive results but generally remained below RoBERTa and FinBERT.

- model families, Machine Learning models consistently outperformed Deep Learning and Attention-based architectures. LightGBM, CatBoost, Random Forest, and SVM repeatedly achieved the highest scores, demonstrating that the contextual embeddings extracted by the fine-tuned transformers already provide highly discriminative representations for sentiment classification.

- Fine-Tuning (FT) consistently outperformed Feature Extraction in the concatenation setting, indicating that adapting the pretrained embeddings to the financial sentiment task captures richer and more task-specific semantic information. This adaptation allows classifiers to exploit more discriminative representations, leading to stronger performance despite the relatively small size of the financial datasets

- While Fine-Tuning achieves the highest absolute scores in the concatenation setting, the benefit brought by concatenation itself is more apparent in the Feature Extraction setting, as it better preserves the complementary information contained in different embedding spaces. During fine-tuning, part of this diversity may be lost through parameter adaptation, resulting in a less pronounced improvement from concatenation, particularly on challenging datasets such as Twitter.

  Overall, the findings suggest that combining two contextual embedding spaces (768+768) is more effective than combining contextual embeddings with the 27-dimensional emotion representation (768+27). While the emotion-based features contribute useful affective information, the richer semantic representations provided by the 768-dimensional embeddings yield superior classification performance. Among all evaluated configurations, RoBERTa-based embeddings achieved the best overall results, with FinBERT remaining the closest competitor.

3.7 Global Study Discussion

3.7.1 Emotion-Sentiment Interaction Analysis

Before evaluating the computational performance of the developed architectures, it is methodologically essential to analyze the underlying emotional structure of the dataset. This subsection investigates the distribution of fine-grained emotions and their direct interaction with high-level financial sentiments.

A. **Global Emotion Distribution**
   To establish a baseline for the fine-grained affective landscape, continuous emotion scores derived from the RoBERTa model fine-tuned on the GoEmotions classification task are aggregated and normalized across four distinct financial validation datasets, and unified across four financial validation datasets used for validation.

   Financial PhraseBank 100% Dataset: As illustrated in Figure 3.9, a uniform structural property emerges in which financial discourse is primarily anchored by the Approval emotion, which commands the highest mean normalized probability. A sharp decline is observed from Approval to Disapproval, followed by Optimism and Realization. This hierarchy suggests that absolute human consensus isolates clear and binary performance evaluations, resulting in highly distinguishable emotional patterns.

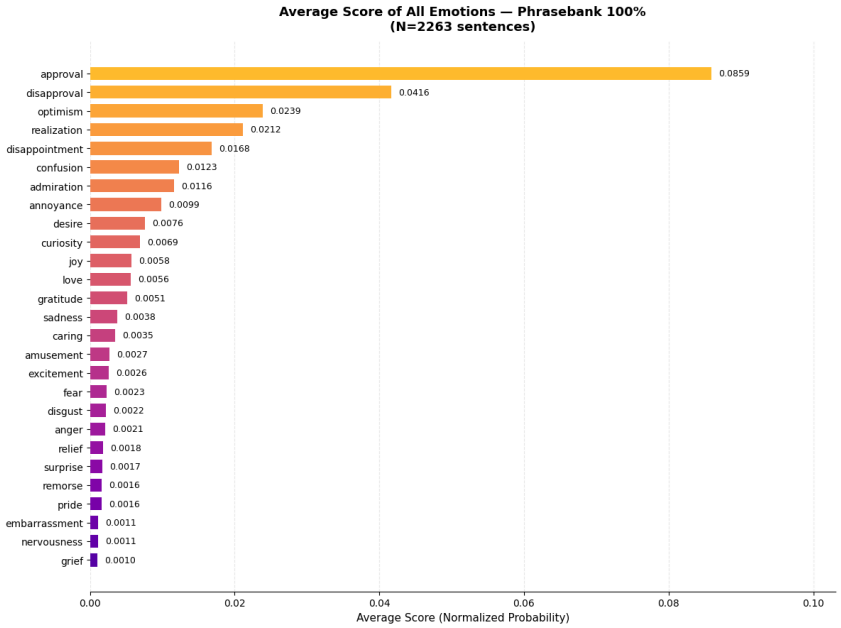


Figure 3.9. Distribution of average emotion scores detected in Phrasebank 100% Dataset

Financial PhraseBank 50% Dataset : As shown in Figure 3.10, Approval remains the dominant emotion. However, Disappointment emerges as the second most prominent emotion, accompanied by a noticeable increase in Sadness. This distribution indicates that texts lacking absolute human consensus contain more nuanced, passive, or soft-toned negative language, reflecting a greater degree of emotional ambiguity.

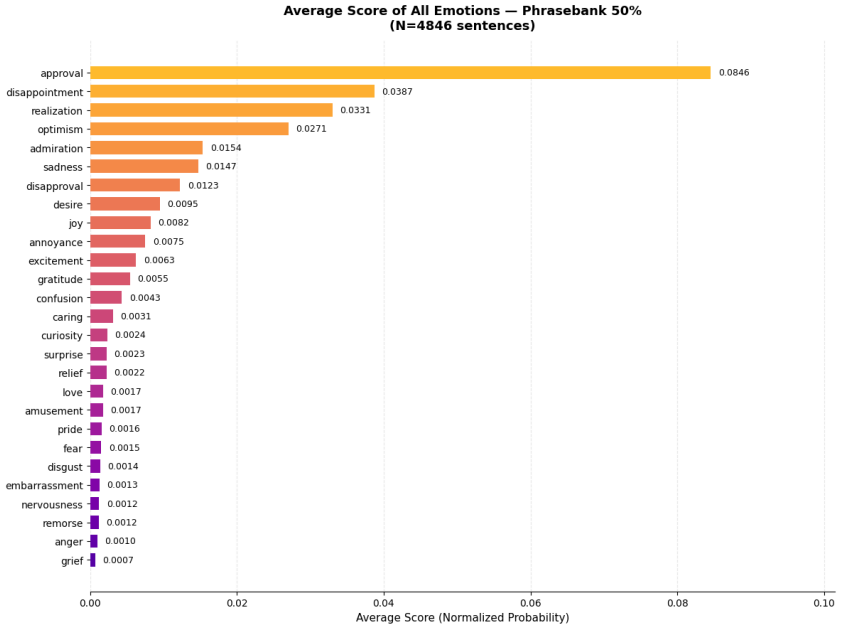


Figure 3.10. Distribution of average emotion scores detected in Phrasebank 50% Dataset

Gold Dataset : Figure 3.11 reveals a distinct rise in Realization and Optimism, both of which surpass Disapproval. This variation reflects the forward-looking nature of analytical commentaries and speculative market insights contained within the Gold dataset. The emotional structure therefore emphasizes interpretation and expectation rather than direct positive–negative opposition.

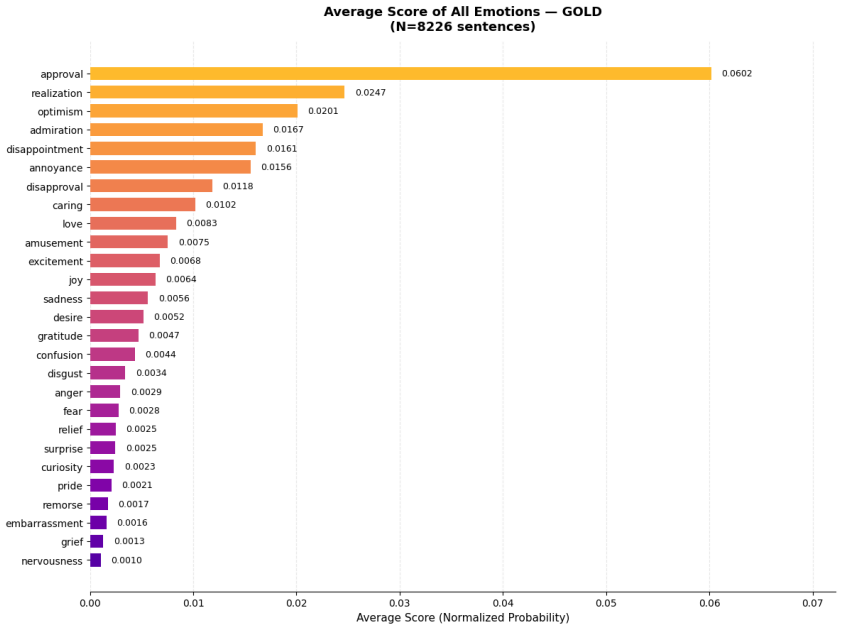


Figure 3.11. Distribution of average emotion scores detected in Gold Dataset

Twitter Dataset : As illustrated in Figure 3.12, Approval remains dominant, followed closely by Disappointment and Curiosity. This distribution highlights the noisy and opinionated nature of social media discourse, where personal reactions, uncertainty, and exploratory attitudes are more prevalent than in professionally curated financial texts.

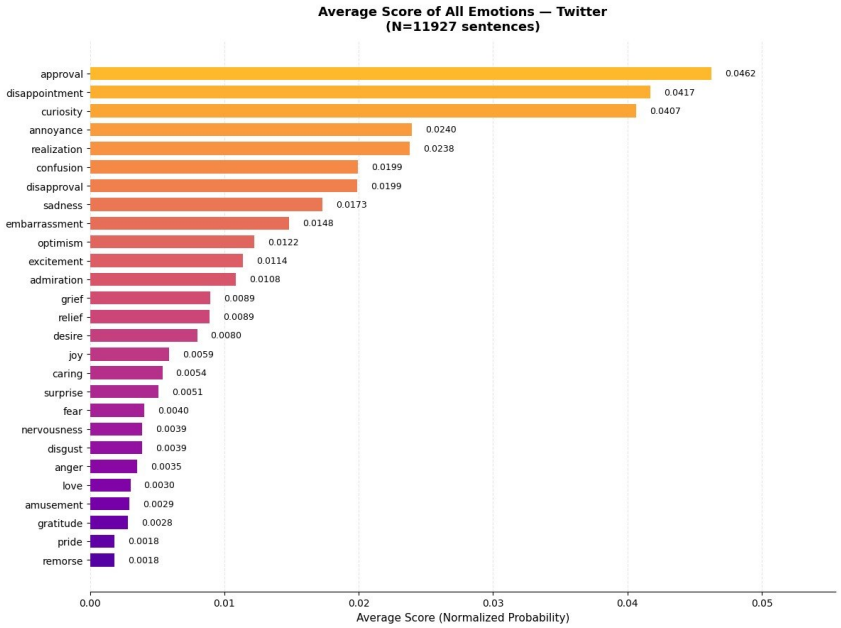


Figure 3.12. Distribution of average emotion scores detected in Twitter Dataset

Across all datasets, intense emotional states such as Nervousness, Grief, and Anger remain negligible, highlighting a persistent class imbalance in which financial discourse suppresses strong affective reactions in favor of analytical and performance-oriented evaluations. Despite the consistent dominance of Approval, the variations observed in secondary emotional categories demonstrate that each dataset possesses a distinct emotional signature. These findings directly support the use of machine learning, deep learning, and attention mechanisms, which are designed to dynamically capture subtle and shifting semantic signals across heterogeneous financial text sources.

B. Dominant Emotions per Sentiment Class

The per-class emotion distributions reveal meaningful affective profiles across sentiment classes and datasets, yielding several theoretically significant observations.

Financial PhraseBank 100% Dataset: Approval strongly dominates all sentiment classes. In the negative class, it leads with 158 sentences (52.1%), far ahead of disappointment (50 sentences, 16.5%). In the neutral class, approval dominates even more clearly (745 sentences, 53.6%), followed by disapproval (411 sentences, 29.6%), reflecting its informational nature. In the positive class, approval also ranks first (251 sentences, 44.0%), followed by disapproval (145 sentences, 25.4%). Intense emotions (anger, fear, sadness, grief) are almost absent across all classes

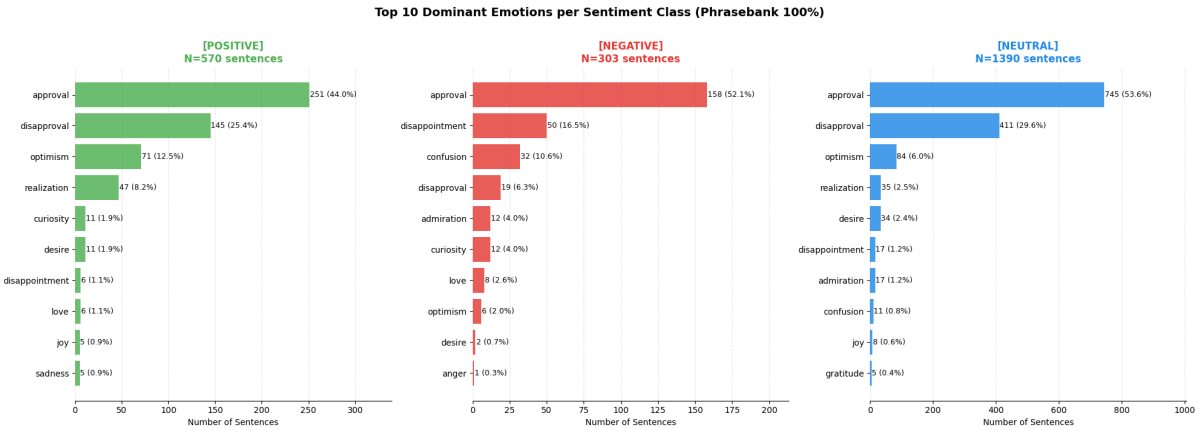


Figure 3.13. Top 10 Emotions by Sentiment Class in the Phrasebank 100% Dataset

Financial PhraseBank 50% Dataset: Approval strongly dominates the positive class with 698 sentences (51.2%), followed by realization (328 sentences, 24.1%). In the negative class, disappointment leads clearly with 291 sentences (48.2%), followed by realization (136 sentences, 22.5%). In the neutral class, approval remains the most dominant (1,166 sentences, 40.5%), but anger appears strongly in second position (700 sentences, 24.3%), followed by realization (610 sentences, 21.2%). Compared to the 100% version, the 50% subset shows a more balanced emotional distribution with higher presence of anger and realization, particularly in neutral and negative sentences.

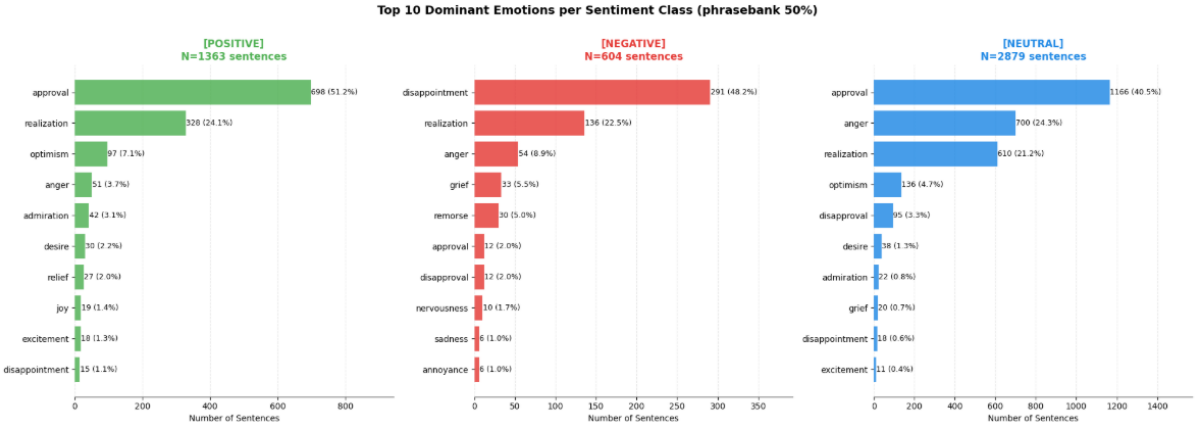


Figure 3.14. Top 10 Emotions by Sentiment Class in the Phrasebank 50% Dataset

Gold Dataset Approval strongly dominates both sentiment classes in the GOLD dataset. In the positive class (N=4,412), it leads overwhelmingly with 2,640 sentences (59.8%), followed by realization (412 sentences, 9.3%) and optimism (314 sentences, 7.1%). In the negative class (N=3,814), approval remains the dominant emotion with 1,630 sentences (42.7%), ahead of realization (281 sentences, 7.4%), optimism (268 sentences, 7.0%), and disappointment (258 sentences, 6.8%). Intense negative emotions such as anger, fear, or sadness are almost absent as dominant emotions. The high dominance of approval across both classes highlights the model's strong bias toward approval-related framing in financial and economic discourse, even in negative contexts.

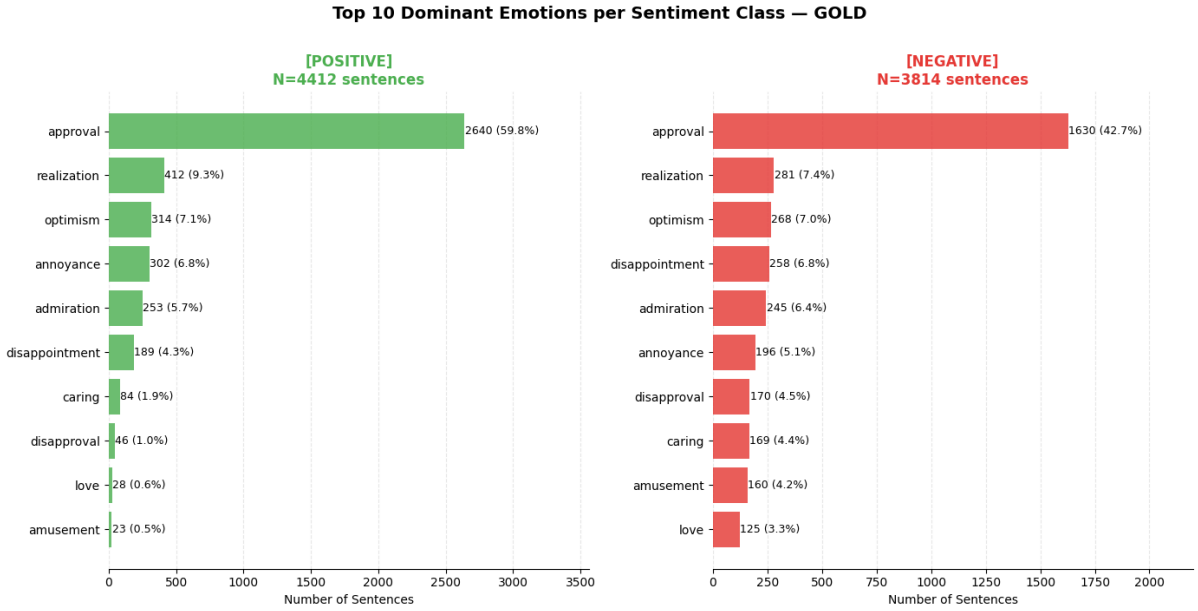


Figure 3.15. Top 10 Emotions by Sentiment Class in the GOLD Dataset

Twitter Dataset: Approval strongly dominates the positive (4,142 sentences, 53.8%) and neutral classes (1,666 sentences, 69.5%). In the negative class, disappointment leads clearly with 754 sentences (42.1%), followed by approval (337 sentences, 18.8%) and annoyance (266 sentences, 14.9%). This distribution shows a more diverse and emotionally expressive profile on Twitter compared to other financial datasets (such as Financial Phrasebank and GOLD), with higher presence of annoyance, curiosity, excitement, and realization, while intense negative emotions remain relatively limited.

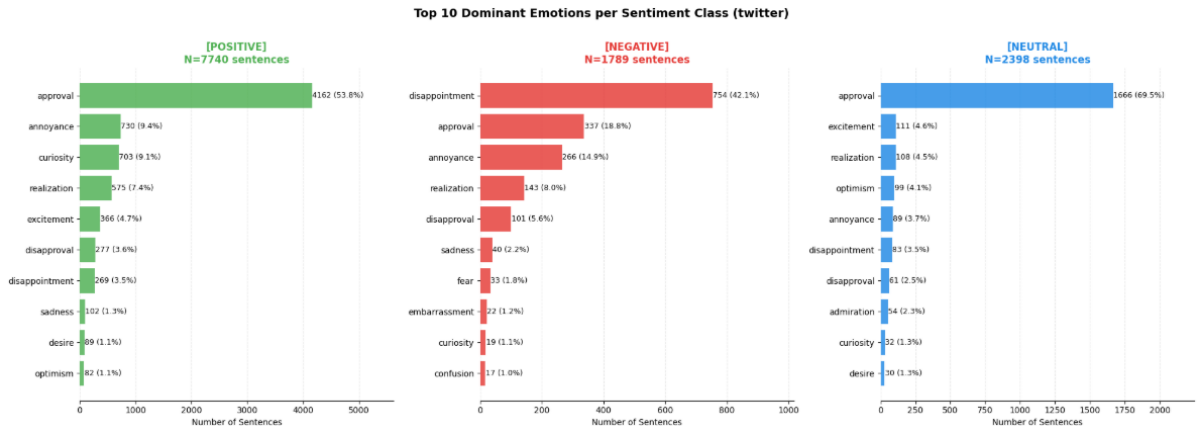


Figure 3.16. Top 10 Emotions by Sentiment Class in the Twitter Dataset

Overall Observation:

Across all datasets, Table 5, which summarizes the top 5 dominant emotions per sentiment class based on the number of sentences where each emotion is the strongest, reveals a remarkably consistent pattern. Approval overwhelmingly dominates across the majority of datasets and sentiment classes, particularly in positive (ranging from 44.0% to 59.8%) and neutral sentences (up to 69.5% in Twitter).

In negative classes, a clear distinction emerges: while Approval remains surprisingly dominant in Phrasebank 100% and GOLD, Disappointment becomes the leading emotion in Phrasebank 50% (48.2%) and Twitter (42.1%). This counterintuitive dominance of Approval in negative financial texts highlights how models tend to interpret factual negative reporting through an evaluative approval lens rather than pure negative affect.

Intense basic emotions such as anger, fear, grief, and sadness remain marginal or negligible in most datasets, with the notable exception of Anger in Phrasebank 50% neutral class. Financial discourse, even on social media (Twitter), consistently favors evaluative (Approval/Disapproval) and cognitive (Realization) signals over raw emotional expressions.

These findings underscore the inherent complexity of financial sentiment annotation and strongly support the necessity of rich contextual language models such as FinBERT, rather than relying solely on compact emotion-based projections, for more accurate and robust financial sentiment classification.

Table 3.14. Top 5 Dominant Emotions per Sentiment Class across Datasets

Dataset	Emotion	Nbre	Weight	Pos	Neg	Neu
FPB 100%	Approval	1154	51.0%	44.0% (251)	52.1% (158)	53.6% (745)
	Disapproval	575	25.4%	25.4% (145)	6.3% (19)	29.6% (411)
	Optimism	161	7.1%	12.5% (71)	2.0% (6)	6.0% (84)
	Realization	82	3.6%	8.2% (47)	—	2.5% (35)
	Disappointment	73	3.2%	1.1% (6)	16.5% (50)	1.2% (17)
FPB 50%	Approval	1876	38.7%	51.2% (698)	2.0% (12)	40.5% (1166)
	Realization	1074	22.2%	24.1% (328)	22.5% (136)	21.2% (610)
	Anger	805	16.6%	3.7% (51)	8.9% (54)	24.3% (700)
	Disappointment	324	6.7%	1.1% (15)	48.2% (291)	0.6% (18)
	Optimism	233	4.8%	7.1% (97)	—	4.7% (136)
Gold	Approval	4270	51.9%	59.8% (2640)	42.7% (1630)	—
	Realization	693	8.4%	9.3% (412)	7.4% (281)	—
	Optimism	582	7.1%	7.1% (314)	7.0% (268)	—
	Admiration	498	6.1%	5.7% (253)	6.4% (245)	—
	Annoyance	498	6.1%	6.8% (302)	5.1% (196)	—
Twitter	Approval	6165	51.7%	53.8% (4162)	18.8% (337)	69.5% (1666)
	Disappointment	1106	9.3%	3.5% (269)	42.1% (754)	3.5% (83)
	Annoyance	1085	9.1%	9.4% (730)	14.9% (266)	3.7% (89)
	Realization	826	6.9%	7.4% (575)	8.0% (143)	4.5% (108)
	Curiosity	754	6.3%	9.1% (703)	1.1% (19)	1.3% (32)

Nbre = total dominant sentence count Weight = Nbre / dataset total
Pos / Neg / Neu = % (count) of sentences where the emotion is dominant within that class
— = emotion absent from top 5 for this class Gold has no Neutral class

3.7.2 Comparison Between Proposed Approach

A. **Approach 1 VS Approach 2**

The comparison between Approach 1 and Approach 2 illustrates the respective strengths of domain-specific pre-training and fine-tuning in financial sentiment classification.
In the feature extraction setting, FinBERT outperforms RoBERTa embeddings in most cases, confirming the advantage of domain-adapted representations when no additional training is applied. Fine-tuning with RoBERTa as the base model produces clear improvements over frozen FinBERT embeddings, with gradient boosting models (CatBoost and LGBM) reaching 98.61% F1-Weighted and 97.32% MCC on Twitter — a substantial increase over the feature extraction setting. Although Approach 2 with full 768-dimensional RoBERTa embeddings and BiGRU+Attention yields competitive results, the fine-tuning strategy in Approach 1 remains more effective by enabling general-domain models to adapt to the financial sentiment task.
Key observations:

- Fine-tuning yields the largest performance gains on noisy data such as Twitter.
- The results demonstrate that domain-specific pre-training (FinBERT) provides significant advantages in feature extraction, while fine-tuning general-purpose models (RoBERTa) allows them to attain superior or comparable performance.
- Approach 1 generally outperforms Approach 2, as employing either a domain-specific or a fine-tuned general model delivers better performance than using unadapted GoEmotions-based RoBERTa embeddings alone.

B. **Approach 1 VS Approach 3**

Feature Extraction (FE)

- The emotion representation derived from GoEmotions contributes positively to sentiment classification, particularly on Twitter, because GoEmotions is built from social media comments and captures emotional patterns commonly found in tweets.
- Despite this contribution, the 768-dimensional contextual embeddings remain more informative than the 27-dimensional emotion vector. The richer semantic and contextual information contained in the full embeddings explains their consistently stronger performance across datasets.

Fine-Tuning (FT)

- The comparison between 768+27 and 768+768 shows that combining two contextual embedding spaces is generally more effective than combining contextual and emotion features. Fine-tuned contextual embeddings preserve richer task-specific information, leading to better overall classification performance.
- Among the fine-tuned transformers, RoBERTa achieved the strongest and most consistent results, while FinBERT remained a close competitor. This suggests that both general-domain knowledge and financial-domain adaptation are valuable, although RoBERTa's representations appear slightly more transferable after fine-tuning.

The results show that strong financial sentiment classification performance can already be achieved using feature extraction alone, without the need for additional fine-tuning. This highlights the effectiveness of pre-trained transformer embeddings in capturing rich semantic and sentiment information. In particular, domain-adapted representations such as FinBERT provide robust features that enable traditional ML, DL, and attention-based classifiers to achieve highly competitive results.Considering all evaluated approaches,

datasets, and experimental configurations, Machine Learning classifiers, particularly LightGBM and CatBoost, consistently achieve the best or near-best performance compared to Deep Learning and Attention-based architectures.

3.7.3 Error Analysis

To better understand model limitations and evaluate the impact of fine-grained emotion modeling, an error analysis was conducted on two selected datasets: the Gold dataset (chosen for its high-quality annotations and formal language) and Financial PhraseBank 50% (selected for its realistic label ambiguity reflecting annotator disagreement). PhraseBank 100% was excluded due to its exceptionally high baseline accuracy leaving little room for measurable improvement, and the Twitter dataset due to excessive noise and class imbalance, respectively.
The primary objective of this analysis is to track the evolution of misclassifications between methodologies. Specifically, we examine the transition from Approach 1, which uses traditional semantic representations via FinBERT feature extraction, to Approach 3, which introduces an enhancement by fusing these semantic layers with fine-grained emotional features using the best-performing model. Feature Extraction (FE) is selected over Fine-Tuning (FT) because it increases the interpretability of the contribution introduced by emotional features. Unlike fine-tuning, which modifies the internal layers of transformer models and may obscure the effect of added features, feature extraction keeps the original semantic embeddings fixed, allowing a clearer measurement of the impact of the introduced emotional cues. This setup enables a direct comparison of confusion matrices to identify which types of financial ambiguities are addressed through the incorporation of structural emotional information.

A. **Gold Dataset**
An error analysis was performed on the Gold dataset using the confusion matrices of the best model in each approach.
Approach 1 (Best: SVM)): Achieved 94.59% Accuracy and 89.13% MCC with 89 misclassifications. The model produced more false positives (46) than false negatives (43), indicating a tendency to classify some negative texts as positive.These misclassifications likely occur on ambiguous or subtle financial texts where sentiment is expressed indirectly (e.g., cautious optimism or mild criticism), which remains challenging even for strong traditional models.

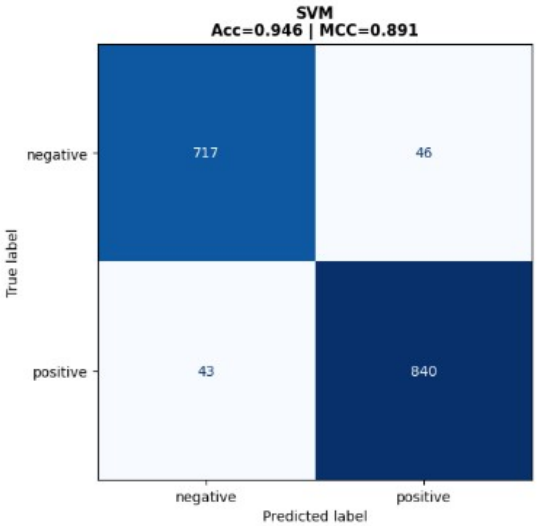


Figure 3.17. Confusion Matrix of SVM (Best Model) in Approach 1 — Gold Dataset

Approach 2 (Best: SVM with 768-dim RoBERTa embeddings):Recorded 94.22% Accuracy and 88.39 MCC. Slightly higher False Positives (54) compared to Approach

1 and misclassifications (95 errors).The model tends to over-predict the positive class on certain negative or neutral financial statements. This behavior may stem from RoBERTa's general-domain pre-training on the GoEmotions dataset, which is less specialized in financial nuance than FinBERT. As a result, the model exhibits higher sensitivity to positive lexical cues, even in contexts where the overall financial sentiment is neutral or slightly negative.

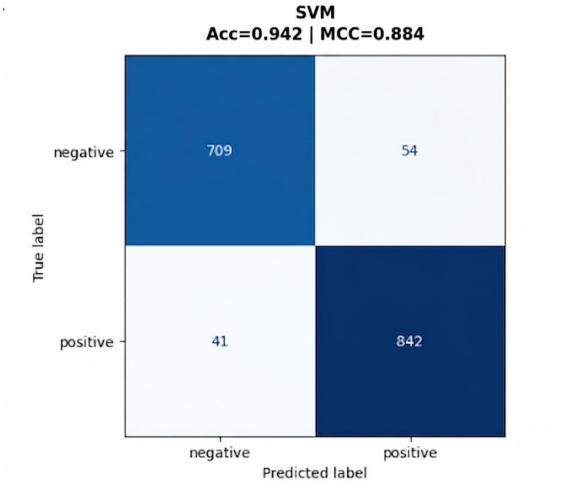


Figure 3.18. Confusion Matrix of SVM (Best Model) in Approach 2 — Gold Dataset

Approach 3 (Best: Self-Attention): delivered the best performance with 95.20% Accuracy and 90.35% MCC and only 79 misclassifications (39 false positives and 40 false negatives).By combining traditional sentiment signals with fine-grained emotion features, the Self-Attention mechanism better captures subtle emotional cues, resulting in fewer confusions between positive and negative classes. The errors that remain are likely due to highly ambiguous texts where even emotional signals do not provide sufficient discriminatory power.

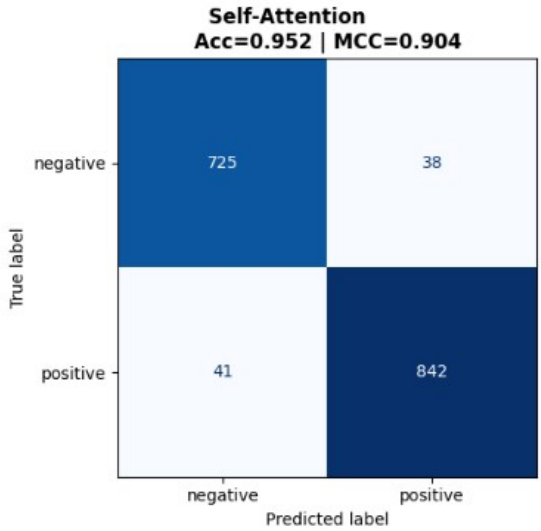


Figure 3.19. Confusion Matrix of Self-Attention (Best Model) in Approach 3 — Gold Dataset

Errors across all approaches mainly occur on borderline cases with implicit or mixed sentiment. Approach 3 achieves the best error distribution, confirming that combining traditional sentiment with fine-grained emotion modeling improves robustness, though highly ambiguous financial texts remain challenging even for the strongest models.

B. **Phrasebank 50% Dataset**

An error analysis was performed on the Phrasebank 50% dataset using the confusion matrices of the best model in each approach.

Approach 1 (Best: XGBoost): Achieved 89.59% Accuracy and 81.32% MCC. The model shows a relatively high number of misclassifications(101 errors), primarily involving the neutral class. It tends to confuse neutral texts with both negative and positive classes, reflecting the inherent ambiguity in texts with only majority agreement. Most errors occur when subtle or mixed financial language is present.

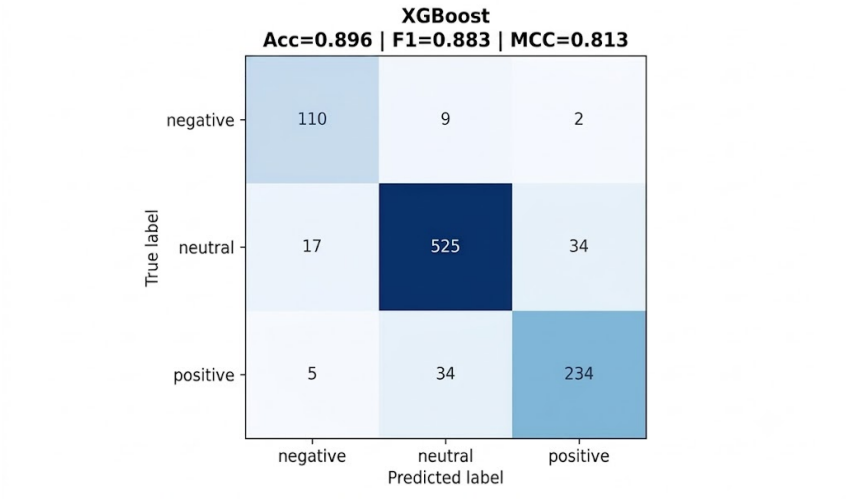


Figure 3.20. Confusion Matrix of XGBoost (Best Model) in Approach 1 — Phrasebank 50% Dataset

Approach 2 (Best: SVM with 768-dim RoBERTa embeddings): Recorded 84.12% Accuracy and 71.34% MCC. This approach has the highest error rate among the three(157 errors). The model struggles significantly with neutral texts, producing a large number of false positives and false negatives around the neutral class. The lower performance may be attributed to RoBERTa's general-domain pre-training on GoEmotions, which is less effective at capturing financial nuances when label noise is present.

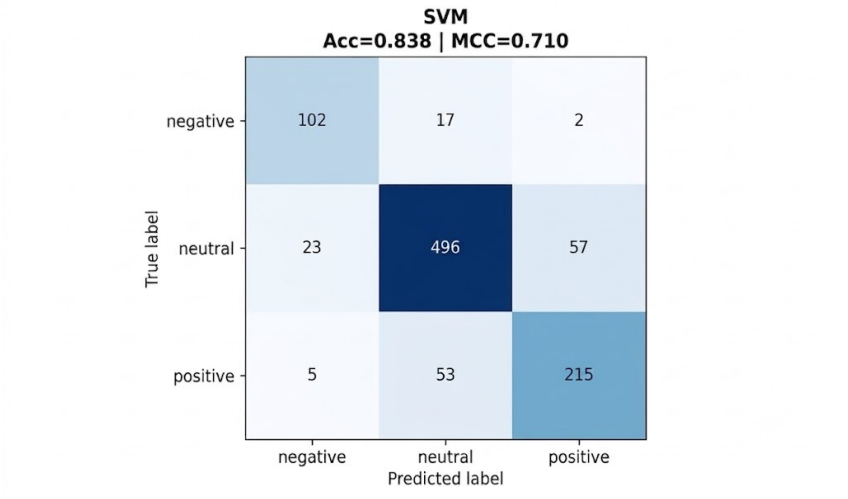


Figure 3.21. Confusion Matrix of -SVM (Best Model) in Approach 2 — Phrasebank 50% Dataset

Approach 3 (Best: SVM): delivered the best performance with 90.9% Accuracy and 82.7% MCC. This hybrid model shows the lowest number of misclassifications (69 errors). By combining traditional sentiment signals with fine-grained emotion features, it better handles the uncertainty of the neutral class and reduces confusion between the three sentiment categories.

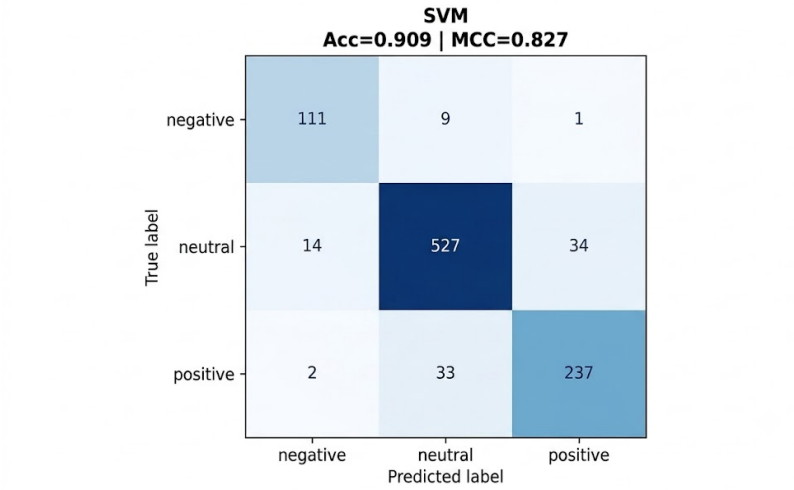


Figure 3.22. Confusion Matrix of SVM (Best Model) in Approach 3 — Phrasebank 50% Dataset

Errors across all approaches are concentrated on the neutral class, which is expected given the majority-agreement nature of Phrasebank 50%. Approach 3 consistently achieves the best error distribution, confirming that integrating fine-grained emotion modeling improves robustness in the presence of label uncertainty. However, even the best models still face difficulties with ambiguous and mixed-sentiment financial texts.

3.8 Comparative With Literature Review

This study's results complement the work of (Mir et al., 2025) in "Joint Topic-Emotion Modeling in Financial Texts: A Novel Approach to Investor Sentiment and Market Trends." Mir et al. present a joint modeling method for both the topics and emotions of investors, thus allowing for more informative and interpretive analysis of financial texts. They report strong performance with 96.3% accuracy and 97.8% F1-score using FinBERT combined with NVDM for joint topic-emotion modeling. Table 3.15 presents a direct comparison between their reported results and those obtained in the present study across multiple datasets.

Table 3.15. Comparison with (Mir et al., 2025) across Datasets

Study	Dataset	Approach	Acc	F1-S	F1-M	F1-W	MCC
(Mir et al., 2025)	StockEmotions	FinBERT + NVDM	96.30	97.80	—	—	—
Present Study (Approach 3)	PhraseBank 100%	FT-RoBERTa + GoEmo (LGBM)	99.55	—	99.58	99.55	99.17
	PhraseBank 50%	FT-RoBERTa + GoEmo (XGB)	92.87	—	91.82	92.85	87.01
	Gold	FT-RoBERTa + GoEmo (BiLSTM)	96.59	—	96.57	96.59	93.15
	Twitter	FT-RoBERTa + GoEmo (LGBM)	92.96	—	90.75	92.92	86.26

The present study adopts a different perspective, focusing on the hybrid fusion of fine-tuned contextual financial representations with fine-grained emotion embeddings (Approach 3). The fusion of fine-tuned RoBERTa with RoBERTa-GoEmotion embeddings achieved the strongest results across all four benchmark datasets, outperforming (Mir et al., 2025) on most metrics as shown in Table 3.15. Specifically, this fusion reached 99.55% F1-W (MCC: 99.17%) on PhraseBank 100% using LightGBM, 92.85% F1-W (MCC: 87.01%) on PhraseBank 50% using XGBoost, 96.59% F1-W (MCC: 93.15%) on Gold using BiLSTM, and 92.92% F1-W (MCC: 86.26%) on the more challenging Twitter dataset using LightGBM. These results consistently surpass the accuracy and F1-score reported by (Mir et al., 2025) on StockEmotions (96.30% accuracy, 97.80% F1-score) using FinBERT combined with NVDM, confirming that combining fine-tuned RoBERTa representations with fine-grained emotion embeddings provides a more robust and generalizable approach for financial sentiment classification.

It is important to note that this comparison should be interpreted as indicative, since the studies differ in terms of datasets, objectives, experimental setups, and evaluation methodologies, which may affect the direct comparability of the reported results.

Overall, this research complements (Mir et al., 2025) by confirming that hybrid feature fusion combining fine-tuned contextual financial embeddings with fine-grained emotion representations can achieve state-of-the-art or superior performance on financial sentiment classification benchmarks, with strong robustness across both well-annotated corpora and noisier social media text.

3.9 Conclusion

This chapter presented the experimental evaluation of the three proposed approaches for financial sentiment classification. The results demonstrated that fine-tuning strategies significantly improve performance compared to feature extraction, particularly on noisy data. The hybrid Approach 3, which combines traditional sentiment analysis with fine-grained emotion modeling, consistently achieved the best overall results across the datasets.

Error analysis on the Gold and Phrasebank 50% datasets confirmed that integrating fine-grained emotions reduces misclassifications on ambiguous and mixed-sentiment texts. Additionally, using full 768-dimensional embeddings proved substantially more effective than reduced representations.

Overall, this study validates that incorporating fine-grained emotion modeling enhances the accuracy and robustness of financial sentiment classification. These findings open promising directions for more nuanced and context-aware financial text analysis systems.

General Conclusion

This study investigated the contribution of fine-grained emotion modeling to financial sentiment analysis through a proposed framework that integrates three complementary approaches: traditional financial sentiment representations, emotion-based representations, and hybrid feature fusion strategies. The experimental results across four benchmark datasets (Financial PhraseBank 100%, Financial PhraseBank 50%, Twitter Financial Sentiment, and Gold Standard) demonstrated that deep semantic representations from pre-trained language models remain the most effective source of information, particularly when using 768-dimensional vectors. Fine-tuning consistently improved performance compared with feature extraction, highlighting the importance of task-specific adaptation for financial sentiment classification.

The findings also showed that emotion information can provide additional value, especially for challenging datasets such as Twitter, where emotional cues are more strongly expressed in social media language. However, the 27-dimensional emotion representation alone was generally insufficient to match the performance of contextual embeddings. In the hybrid framework, combining emotion-based and financial representations produced competitive results, although the fusion of high-dimensional contextual embeddings (768+768) consistently outperformed combinations involving the 27-dimensional emotion vector (27+768). This indicates that richer semantic representations preserve more sentiment-relevant information than compressed emotion distributions.

Among the evaluated transformer models, RoBERTa achieved the strongest and most consistent results after fine-tuning, while FinBERT remained highly competitive due to its financial-domain specialization. Specifically, the fusion of fine-tuned RoBERTa with RoBERTa-GoEmotion embeddings achieved the best overall results, reaching an F1-Weighted score of 99.55% on PhraseBank 100% using LightGBM, 92.85% on PhraseBank 50% using XGBoost, and 96.59% on the Gold dataset using BiLSTM. Twitter remained the most challenging benchmark due to its informal and noisy language; nevertheless, the fine-tuned RoBERTa fusion achieved the best performance on this dataset as well, with an F1-Weighted score of 92.92% using LightGBM.

For future work, larger financial language models, advanced feature fusion mechanisms, and attention-based multimodal architectures could be explored to better integrate emotion and sentiment information. Additionally, extending the evaluation to multilingual financial datasets and real-time market data may provide further insights into the relationship between emotions and financial sentiment. Overall, the study confirms that fine-tuned contextual embeddings constitute the most effective solution for financial sentiment analysis, while fine-grained emotions offer complementary information that is particularly beneficial in social media environments.

Bibliography

Adams, T., Ajello, A., Silva, D., & Vazquez-Grande, F. (2023). More than words: Twitter chatter and financial market sentiment. *arXiv preprint arXiv:2305.16164*.

Aditya, M. R., Sutanto, T., Budiman, H., Ridha, M. R. N., Syapotro, U., & Azijah, N. (2024). Machine learning models for classification of anemia from cbc results: Random forest, svm, and logistic regression. *Journal of Data Science*, *2024*.

Alkaabi, H., Jasim, A. K., & Darroudi, A. (2025). From static to contextual: A survey of embedding advances in nlp. *PERFECT: Journal of Smart Algorithms*, *2*(2), 57–66.

Alqarni, F., Sagheer, A., Alabbad, A., & Hamdoun, H. (2025). Emotion-aware roberta enhanced with emotion-specific attention and tf-idf gating for fine-grained emotion recognition. *Scientific Reports*, *15*(1), 17617.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Cheng, W., & Chen, S. (2021). Sentiment analysis of financial texts based on attention mechanism of finbert and bilstm. *2021 International Conference on Computer Engineering and Application (ICCEA)*, 73–78.

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 6.

Choi, S. R., & Lee, M. (2023). Transformer architecture and attention mechanisms in genome data analysis: A comprehensive review. *Biology*, *12*(7), 1033.

Cicekyurt, E., & Bakal, G. (2025). Enhancing sentiment analysis in stock market tweets through bert-based knowledge transfer. *Computational Economics*, 1–23.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th annual meeting of the association for computational linguistics*, 8440–8451.

Costa, R. (2026). Twitter financial news sentiment dataset. %5Curl%7Bhttps://www.kaggle.com/datasets/sulphatet/twitter-financial-news%7D

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. *Proceedings of the 58th annual meeting of the association for computational linguistics*, 4040–4054.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Du, K., Xing, F., Mao, R., & Cambria, E. (2024). Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, *56*(9), 1–42.

Gianstefani, I., Longo, L., & Riccaboni, M. (2022). The echo chamber effect resounds on financial markets: A social media alert system for meme stocks. arxiv. org.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: An overview. *arXiv preprint arXiv:2008.05756*.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hoa, P. V., Binh, N. A., Hong, P. V., An, N. N., Thao, G. T. P., Hanh, N. C., Ngo, P. T. T., & Bui, D. T. (2024). One-dimensional deep learning driven geospatial analysis for flash flood susceptibility mapping: A case study in north central vietnam. *Earth Science Informatics*, *17*(5), 4419–4440.

Jatavallabha, A., Gerlach, J., & Naresh, A. (2024). Deciphering air travel disruptions: A machine learning approach. *arXiv preprint arXiv:2408.02802*.

Kalyani, J., Bharathi, P., Jyothi, P., et al. (2016). Stock trend prediction using news sentiment analysis. *arXiv preprint arXiv:1607.01958*.

Khan, M. Y., Qayoom, A., Nizami, M. S., Siddiqui, M. S., Wasi, S., & Raazi, S. M. K.-R. (2021). Automated prediction of good dictionary examples (gdex): A comprehensive experiment with distant supervision, machine learning, and word embedding-based deep learning techniques. *Complexity*, *2021*(1), 2553199.

Kocoń, J., Gruza, M., Bielaniewicz, J., Grimling, D., Kanclerz, K., Miłkowski, P., & Kazienko, P. (2021). Learning personal human biases and representations for subjective tasks in natural language processing. *2021 IEEE international conference on data mining (ICDM)*, 1168–1173.

Li, X., Al Ansari, N., & Kaufman, A. (2025). Stocktwits: Comprehensive records of a financial social media platform from 2008 to 2022. *Journal of Quantitative Description: Digital Media*, *5*.

Liang, W., Li, Z., & Chen, W. (2024). Enhancing financial market predictions: Causality-driven feature selection. *International Conference on Advanced Data Mining and Applications*, 149–163.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lokker, C., Abdelkader, W., Bagheri, E., Parrish, R., Cotoi, C., Navarro, T., Germini, F., Linkins, L.-A., Haynes, R. B., Chu, L., et al. (2024). Boosting efficiency in a clinical literature surveillance system with lightgbm. *PLOS digital health*, *3*(9), e0000299.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, *66*(1), 35–65.

Lowe, S. (2026). Roberta-base-go_emotions. https://huggingface.co/SamLowe/roberta-base-go_emotions

Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., & Balahur, A. (2018). Www'18 open challenge: Financial opinion mining and question answering. *Companion proceedings of the the web conference 2018*, 1941–1942.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, *65*(4), 782–796.

Mazari, A. C., Benterkia, A., & Takdenti, Z. (2024). Advancing offensive language detection in arabic social media: A bert-based ensemble learning approach. *Social Network Analysis and Mining*, *14*(1), 186.

Mazari, A. C., Boudoukhani, N., & Djeffal, A. (2024). Bert-based ensemble learning for multi-aspect hate speech detection. *Cluster Computing*, *27*(1), 325–339.

Mazari, A. C., & Djeffal, A. (2021). Deep learning-based sentiment analysis of algerian dialect during hirak 2019. *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*, 233–236.

Mazari, A. C., & Djeffal, A. (2022). Sentiment analysis of algerian dialect using machine learning and deep learning with word2vec. *Informatica*, *46*(6).

Mazari, A. C., & Kheddar, H. (2023). Deep learning-based analysis of algerian dialect dataset targeted hate speech, offensive language and cyberbullying. *International Journal of Computing and Digital Systems*.

McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*(1), 51–56.

Mir, M. N. H., Bhuiyan, M. S. M., Al Rafi, M., Rodrigues, G. N., Mridha, M. F., Hamid, M. A., Monowar, M. M., & Uddin, M. Z. (2025). Joint topic-emotion modeling in financial texts: A novel approach to investor sentiment and market trends. *IEEE Access*.

Ngo, A., & Kocoń, J. (2025). Integrating personalized and contextual information in fine-grained emotion recognition in text: A multi-source fusion approach with explainability. *Information Fusion*, *118*, 102966.

Park, E. H., & Storey, V. C. (2023). Emotion ontology studies: A framework for expressing feelings digitally and its application to sentiment analysis. *ACM Computing Surveys*, *55*(9), 1–38.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Rajapaksha, P., Farahbakhsh, R., & Crespi, N. (2021). Bert, xlnet or roberta: The best transfer learning model to detect clickbaits. *IEEE Access*, *9*, 154704–154716.

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. *Proceedings of the 57th annual meeting of the association for computational linguistics*, 5370–5381.

Ren, Z., Yang, Z., Ye, C., Sun, H., Chen, C., Zhu, X., & Liao, X. (2025). Fine-grained emotion recognition via in-context learning. *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 2503–2513.

Rivera-Zavala, R. M., & Martínez, P. (2021). Analyzing transfer learning impact in biomedical cross-lingual named entity recognition and normalization. *BMC bioinformatics*, *22*(Suppl 1), 601.

Roldan, A. (2023). Sentiment analysis in the financial phrasebank dataset: An approach to finbert. *Res. Roldanresearchgate.net*.

Saraswat, P., & Raj, S. (2021). A brief review on machine learning and its various techniques. *International Journal of Innovative Research in Computer Science and Technology*, *9*(6), 110–113. https://doi.org/10.55524/ijircst.2021.9.6.25

Shi, L., Jia, Z., & Liu, R. (2025). Expert-level financial sentiment analysis via internal and external financial knowledge fusion. *Information Fusion*, 103884.

Shiri, F. M., Perumal, T., Mustapha, N., & Mohamed, R. (2023). A comprehensive overview and comparative analysis on deep learning models: Cnn, rnn, lstm, gru. *arXiv preprint arXiv:2305.17473*.

Sinha, A., & Khandait, T. (2021). Impact of news on the commodity market: Dataset and results. *Future of Information and Communication Conference*, 589–601.

Sun, Y., Yuan, H., & Xu, F. (2025). Financial sentiment analysis for pre-trained language models incorporating dictionary knowledge and neutral features. *Natural Language Processing Journal*, *11*, 100148.

Suresh, V., & Ong, D. C. (2021). Using knowledge-embedded attention to augment pre-trained language models for fine-grained emotion recognition. *arXiv preprint arXiv:2108.00194*.

Tomar, S., Dembla, D., & Chaba, Y. (2024). Analysis and enhancement of prediction of cardiovascular disease diagnosis using machine learning models svm, sgd, and xgboost. *International Journal of Advanced Computer Science & Applications*, *15*(4).

Tong, J., Li, S., Zhuang, Z., Hu, J., & Zou, Y. (2025). Emosync: Multi-stage reasoning with multimodal large language models for fine-grained emotion recognition. *Proceedings of the 3rd International Workshop on Multimodal and Responsible Affective Computing*, 95–99.

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. *Artificial intelligence review*, *53*(8), 5929–5955.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, K., Jing, Z., Su, Y., & Han, Y. (2024). Large language models on fine-grained emotion detection dataset with data augmentation and transfer learning. *arXiv preprint arXiv:2403.06108*.

Wang, R., Xu, D., Cascone, L., Wang, Y., Chen, H., Zheng, J., & Zhu, X. (2025). Raft: Robust adversarial fusion transformer for multimodal sentiment analysis. *Array*, 100445.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1480–1489.

Zhang, M., Chen, J., & Palade, V. (2024). Fine-grained sentiment analysis for enhanced financial distress prediction. *2024 IEEE 3rd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, 170–180.

Zhong, P., Wang, D., & Miao, C. (2019). Knowledge-enriched transformer for emotion detection in textual conversations. *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 165–176.