Application of AI in Financial Sector: Earnings Call Dataset Analysis

Kesa M. Abbas

Application of AI in Financial Sector: Earnings Call Dataset Analysis

Kesa M. Abbas

December 2023

A Thesis Submitted
in Partial Fulfillment
of the Requirements for the Degree of
Master of Science
in
Computer Engineering

RIT | Kate Gleason College of Engineering

Department of Computer Engineering

Application of AI in Financial Sector: Earnings Call Dataset Analysis

Kesa M. Abbas

Committee Approval:

Dongfang Liu *Advisor* Date
Department of Computer Engineering

Michael Zuzak Date
Department of Computer Engineering

Andres Kwasinski Date
Department of Computer Engineering

Acknowledgments

I wish to express my profound gratitude to my research advisor, Dr. DongFang Liu, for his invaluable mentorship, assistance, and support throughout the research and writing of this thesis. My appreciation also extends to my committee members, Dr. Michael Zuzak and Dr. Andres Kwasinski, for their guidance. I would also like to thank Nicholas Curl for his help. Lastly, I am deeply grateful to my family and friends for their unwavering support.

To Mummy, for always believing in and supporting me, even when I doubted myself. To Abbu, for his unwavering love and continuous support. To Khala and Khalu, who have been like a second set of parents to me throughout this journey. To Nanno, who has always been my biggest cheerleader and hype-woman. To my dear friends, Long and Tianran, for their enduring love and steadfast support.

And lastly, to myself, for persevering even in the face of adversity. The playlist recommendation while reading this is Skill Issue.

Abstract

Deep learning has emerged as a cornerstone in diverse scientific domains, demonstrating profound implications both in academic research and conventional applications. The utilization of deep learning has branched to the financial sector for stock prediction, management of assets, credit score analysis, etc. During the end of the financial fiscal year, top executives present their companies' growth and losses in earnings calls. These earnings calls invariably webcasted, furnish stakeholders with insights into a firm's fiscal performance during a given period. These audios along with transcripts are published on the company website for the general public. If these audio and transcripts are collected over time, they are capable of forming an extensive audio and sentiment analysis database, which could be utilized to train a model effectively. The quality and quantity of the dataset are essential for a machine-learning model. The availability of these discussions in both audio and text formats on corporate portals presents an invaluable repository for longitudinal audio and sentiment analysis.

The clarity in the statistics of the earnings calls would also predict the bona fide outlook of the companies' financial prospects. This thesis work will focus on adding and extending the earnings call data for the companies for MAEC dataset and create a diverse, updated machine learning dataset. This would be further followed by the proposal of a sentiment analysis tool inspired by the deep neural network to label the data collected and form the baseline method for future work to compare. The established algorithm would provide a pathway for a dataset and data annotation that would further be utilized by the models out there, to perform financial analysis and factual correctness of the data presented, and draw appropriate conclusions. The addition to this dataset and the creation of an algorithm would involve rigorous empirical investigation and the iterative design of a methodological framework conducive to the systematic accrual and annotation of earnings call records.

Contents

List of Figures

List of Tables

Acronyms

AI

Artificial Intelligence

ANN

Artificial Neural Network

API

Application Program Interface

AUC

Area Under Curve

CLEVR

Compositional Language and Elementary Visual Reasoning

CNN

Convolutional Neural Network

CNN-LSTM

Convolutional Neural Network - Long-Short-Term Memory

COCO

Common Objects In Context

DOM

Document Object Model

FFD

Financial Fraud Detection

FP

False Positve

FPR

False Positive Rate

HNR

Harmonics to Noise Ratio

ILSVRC

ImageNet Large Scale Visual Recognition Challenge

MAEC

A Multimodal Aligned Earnings Conference Call Dataset for Financial Risk Prediction

NHR

Noise to Harmonics Ratio

NLP

Natural Language Processing

ROC

Receiver Operating Characteristic Curve

TP

True Positive

TPR

True Positive Rate

URL

Uniform Resource Locator

VQA

Visual Question Answering

Chapter 1

Introduction

1.1 Motivation

Real-time object tracking [6, 7] has had few challenges in the past, and the inclusion of live sentiment analysis increases the complexity of the model created. There are quality machine learning algorithms that have been selectively trained on the limited datasets currently present. Datasets form the backbone of any machine learning implementations out there. They are made up of images, texts, audio, videos, numerical data points, etc., for performing various artificial intelligent tasks such as categorizing, classifying, sentiment analysis, or predicting image and video algorithms. The datasets are applied to all the practical fields out there such as object detection, facial recognition, image, and video classification, speech, sentiment, and stock market analysis are some of, where the need for a quality dataset is out there [8]. The more data available the better the quality of the machine learning algorithm.

Moreover, the dynamic nature of real-time object tracking coupled with sentiment analysis makes the task not just computationally challenging but also intricate in terms of contextual understanding. The present landscape of machine learning has been more inclined towards models that have been fine-tuned with available datasets. But one must remember that datasets aren't just passive repositories of information; they are reflections of the evolving real world. As such, the datasets need to be

continuously updated, enriched, and diversified to keep pace with the ever-evolving dynamics of our world. The role of curated datasets in determining the efficacy of machine learning models cannot be understated[9]. They are the lens through which algorithms perceive, learn, and subsequently, act. With the exponential increase in data generation, especially in fields like object tracking and sentiment analysis, the next frontier in AI would likely revolve around how efficiently and holistically this data can be harnessed. Emphasizing this, the holistic development and curation of datasets should be a priority for researchers and institutions aiming to make meaningful advancements in AI.

1.2 Datasets and Financial Crashes

The corporate world has been subjected to a series of accounting scandals, such as Enron and World Com occurred in the 2000s, the collapse of Lehman Brothers in the 2008 stock-market crash which resulted in significant damage and loss of trust in the corporate market, and insufficient and inefficient allocation of funds in the stock markets. The lack of analysis of the fraudulent statistics, and data, has resulted in distrust and injudicious investments, thus, hurting the investors and the corporations simultaneously [10]. The machine learning datasets are prone to biases, human errors, privacy and compliance issues, security violations, and incorrect labeling of the data. Thereby, decreasing the reliability and authenticity of models tested and trained on these datasets [11]. The unpredictability of datasets acutely affects the financial machine-learning models which are utilized to detect the threat of corporate financial fraud. Currently, there are limited methods available to identify investor frauds, thus, putting the investors and companies at risk and prone to financial frauds, thus, potentially hurting a large amount of capital involved in these transactions. Investors rely on financial statistics to judge and invest their capital and with the availability of falsified data and fraudulent statistics, they are prone to financial fraud. There has

been a proposed implementation of the utilization of sentimental analysis, audio and visual tracking along with financial statistics to develop tools to detect the fraudulent data presented in the earnings call. But there has been less success in the union of the two fields due to the lack of quantity and quality of the data out there [10].

The necessity for rigorous and comprehensive evaluation tools in the financial sector cannot be overstressed, especially in an era where global economies are becoming intricately interlinked. Any misstep or oversight can ripple across continents, impacting economies, jobs, and lives at an unprecedented scale. The challenges faced by the current analytical tools, stemming primarily from their inherent limitations in capturing the multi-dimensional nature of corporate disclosures, necessitate a revolutionary shift in approach. Beyond the quantitative metrics, the intangible elements, like executive confidence or hesitancy during an earnings call, can provide invaluable insights. Harnessing the combined power of sentiment analysis, audio cues, and traditional financial indicators could usher in a new era of financial scrutiny. By meticulously integrating these diverse data streams, there's an opportunity to develop a more resilient and predictive system, ensuring that stakeholders can make informed decisions based on a rich tapestry of information [12]. Only by evolving and adapting to the nuances of the modern corporate landscape can the financial sector hope to avert crises and maintain the integrity of the global economic framework.

1.3 Objective

The primary focus of this research is to expand and diversify the earnings call statistics dataset. This would be further followed by the development of a sentiment analysis tool, that would categorize the textual data into emotions. This would involve an extensive literature review and subsequent development of the methodology for the collection of the earnings call video and audio transcripts. The development of this scheme would provide a convenient technique to retrieve earnings call data from the

corporations' websites. This would also include the identification of the best route to develop software to perform the collection of data. This earnings call data collected then would be annotated by a model inspired by a deep neural network which would form the baseline for the other models out there. This would involve adding labels, classifications, or other metadata to the data points in the dataset, in order to provide additional context and meaning to the machine learning model. Thereby, leading to the creation of a dataset, which would be utilized by the other models to train and test, and accurately perform real-time object tracking and sentiment analysis.

Chapter 2

Background

2.1 Accounting Scandals

The advent of the stock market crash in 2008, and the fraudulent depiction of the company data prompted the utilization of vocal and linguistic cues for the fraud analysis in the earnings call. The verbal cues, such as stammering, and laughing a specific pitched noise when fraudulent company statistics were mentioned [13]. If these vocal and physical cues were effectively recorded and noted in the live-stream videos, they could be utilized to train a model and perform sentiment and information analysis in real-time. Sentiment analysis, therefore, would play a critical role, in the assessment of the facial cues with the factual presentation of the company statistics, thereby, providing investors protection against a financial scandal[14].

Building on the significance of these discoveries, it's worth noting that the intricacies of human expression, both vocal and physical, can offer invaluable insights into the underpinnings of corporate communications. In the high-stakes world of financial markets, where the slightest hint of deceit or manipulation can lead to vast economic repercussions, harnessing the power of sentiment analysis becomes even more paramount. Not only does it act as a potential safeguard against financial misinformation, but it also offers a promising avenue for investors and stakeholders to glean deeper, more nuanced understandings of company presentations. In the age of tech-

nology and information, the ability to decipher these subtle cues and nuances, and juxtapose them against the hard data presented, could very well be the game-changer in ensuring transparency and trustworthiness in the corporate landscape

2.2 Natural Language Processing for Earnings Call

The development of NLP occurred in the 1950s when the concept of linguistics and machine learning were cross-trained [15]. This approach aided the machines to obtain the ability to learn, read and interpret spoken language and written text while simultaneously generating accompanying statistics from the given interaction. The current boom of artificial intelligence along with the earnings call has given rise to three major setups:

- Development of new algorithms and techniques.
- Optimization and re-inventing the existing machine learning algorithms.
- The availability of large-scale datasets.

Furthermore, the symbiotic relationship between NLP and the exponential growth of data cannot be overstated. As we entered the era of big data, the demand for effective language processing tools skyrocketed. Every byte of data processed, whether it's a tweet, a research paper, or an earnings call transcript, adds to the vast tapestry of linguistic patterns that machines can learn from [15].

Another noteworthy transformation in the NLP landscape is the growing emphasis on cross-linguistic and cross-cultural studies. Recognizing that businesses and research endeavors are no longer confined to linguistic or geographic silos, the tools and techniques developed in the NLP domain are striving for universality [16]. This global approach ensures that machines are equipped to interpret and generate content across diverse linguistic terrains, a feat that holds particular significance in the globalized corporate world.

Moreover, the intersection of NLP with other domains like cognitive science, behavioral finance, and sociolinguistics is leading to more holistic analytical tools. For instance, in the context of earnings calls, it's not just about parsing the language anymore. It's about understanding the cultural, psychological, and sociological undertones of the speaker's utterances [17]. By integrating these multi-disciplinary insights, NLP tools are evolving into more than just language processors; they're morphing into comprehensive human communication analyzers.

The culmination of these advancements promises an exciting future for NLP. As datasets continue to grow, algorithms become more refined, and interdisciplinary collaborations flourish, we stand on the brink of a new horizon in linguistic understanding, one where the boundaries between human intuition and machine intelligence grow increasingly blurred.

2.3 Development and Testing of Current FFD Methods

The conventional FFD methods involve keeping track of the company statistics, the headcount of the employees, stock price, employee reviews, and sudden reductions in the workforce. These commercial methods in the past were substantial to draw the numerical statistical situation of the companies. Currently, the attention is focused towards:

- Linguistic observations
- Vocal observations

Over the course of the period, these have been analyzed separately to draw conclusions about the financial status of the companies. FFD has had some linguistic success.

Humpherys et al. obtained information by analyzing the Management's Discussion and Analysis (MD & A) portion of the Form.10-Ks are yearly reports that public

firms must file with the Securities and Exchange Commission (SEC), [18]. These forms are designed to offer an overview of the financial and business health of the company. Humpherys et al. obtained up to 67% accuracy in identifying fraud by employing language factors such as lexical variety and syntactic complexity. Bloomfield [19] proposed that a linguistic study of CEOs' spontaneous speech in comparison to the legally validated material contained in MD&As might yield further information. Larcker and Zakolyukina [20] conducted a textual study of linguistic traits that recognized deceptive discussions better than chance levels in CEOs' communications with financial analysts during quarterly earnings conference calls.

2.4 Earnings Call with Audio and Text Analysis

Over the course of recent years, the traditional methods have proved short of the current developments with the forging of company statistics, and the practiced ease of acting to review the fraudulent statistics. The earnings conference call are capable of being analyzed to build a multi-modal analysis problem incorporating textual and audio information.

2.4.1 Text-Based Financial Datasets

The financial reports and datasets published for the masses are in textual forms. Over the course of time, it has become significantly difficult to assess the validity and draw relevant statistics from a large amount of randomized data. There is an increasing demand for organized and quality datasets for the appropriate conclusion from the statistics. In recent examinations of 10-K filings and earnings calls, for example, many intriguing findings were supported by transcripts for risk prediction and information from financial disclosure datasets. Past work demonstrates pragmatics and semantics and earnings calls have a huge impact on analysts' decision-making. Analysts evaluate target suggestions based on the advice of the corporation following the quarterly

earnings calls. In fact, [21] indicates in very early research that language-based models might expose misleading information during earnings calls and afterward trigger stock price movements in financial markets. Unlike other text-driven jobs that have rich datasets to exploit, such as financial news datasets presented [22],[23], [24], tweets data used by, and 10-K report datasets released by [15].

Financial text analysis, particularly of earnings calls and financial reports, has opened up a novel paradigm for investors, market analysts, and researchers. Traditional datasets, though voluminous, often miss out on capturing the nuanced linguistic intricacies that reveal deeper insights into a company's financial health and forward-looking perspectives.

The dynamic nature of the corporate world further complicates this scenario. For instance, while some companies might use overly optimistic language to hide underlying issues, others might present a conservative front despite enjoying a robust financial position [25]. The sentiment and tone reflected in these reports often act as indicators, signaling potential corporate strategies and market movements. Recognizing these subtle cues, therefore, can be a game-changer in the world of financial analytics.

Moreover, with the rapid digitization of financial data and the surge of unstructured data, the traditional text-processing methodologies often fall short. The vast array of online forums, analyst commentary, and investor feedback on various platforms provide a gold mine of sentiment indicators. But tapping into this reservoir requires sophisticated NLP tools capable of handling both the volume and complexity of the data [26].

The integration of machine learning and NLP in financial analysis is a promising avenue. Models can be trained to recognize patterns and anomalies, identify sentiments, and even predict market reactions based on historical data. However, this endeavor isn't without its challenges. The sheer variety of financial documents, the

diversity in the writing style across regions, companies, and time periods, and the evolving jargon of the financial world all pose significant hurdles.

Despite these challenges, there's growing recognition of the potential rewards. As cited in the works of [21], [22], and others, there's a profound impact of language patterns on market reactions. Going forward, a combination of robust datasets, cutting-edge NLP tools, and interdisciplinary collaboration will be pivotal in harnessing the full potential of financial text analysis. This is not just about enhancing stock predictions but reshaping the very foundation of financial research and decision-making.

2.4.2 Multi-Modal Financial Datasets

Currently, multi-media processing advancements have been built on combining various sorts of characteristics together during training. Some common datasets have been generated to accommodate Multimodal learning and high-level embedding from many data sources are progressing. The Vision-and-Language BERT (ViBERT) [4], for example, requires pictures and natural language datasets, whereas M-BERT [27] investigates the prospect of leveraging language. This dataset combines audio and visual data based on the transformer architecture. There has always been a focus on the development of a text-audio-aligned dataset, inspired by the statistics and data released by S & P companies at the end of every fiscal year. The S & P 500 Earnings Conference Calls dataset [28] was the first attempt to consider earnings call analysis as a multimodal (text + audio) potential because of the consistent requirement of it in FFD methods to train the models.

Past research has proven that the emotions and psychological attributes of the speaker can be abstracted and a relationship can be drawn between them. Consistent research has demonstrated that various emotions and physical activities of a speaker could be utilized to be abstracted and represented. Consequently, the recent models provide insight into text-aligned data to capture verbal features along with vocal

cues. This aids in the semantic and pragmatic features extracted from the audio transcripts to provide statistics [15]. Li, et, al collected the randomized data released by the companies over the period of three years and sorted and aligned them into a multimodal aligned dataset earning call, considering the textual data. These provide the starting point for the dataset needed for the sentiment analysis and conclusion deduction from the earnings call data. This provides a basis for further research in the analysis of the earnings call [15].

2.5 Related Work

Data is notably important in helping computers comprehend images as collections of items, which humans do naturally but robots have so far struggled to do. More specifically, it would be ideal for computers to classify the objects in an image automatically (image classification), locate them accurately (object detection), and determine which of them are interacting and in what ways (visual relationship detection) [1, 29]. To form the baseline for the research, some popular datasets and their methodology for the creation of their respective datasets would be conferred.

A deep dive into the methods and algorithms that power these tasks reveals a heavy reliance on extensive datasets. These datasets, curated meticulously, are collections of labeled images that serve as the foundation for training machine learning models. The labeling process involves human annotators who assign categories or descriptions to various elements within an image, essentially creating a reference guide for the machine to learn from.

Several renowned datasets have set the standard in the computer vision community, each with its unique set of images, annotations, and use cases. Understanding the methodologies employed in curating these datasets is paramount, as it sheds light on the challenges faced, solutions employed, and the breadth and depth of information captured. By examining the gold standards in the field, researchers can glean

insights for their own projects, ensuring their models are trained on data that is both representative and comprehensive. The rigorous process behind these datasets exemplifies the importance of quality data in pushing the boundaries of what machines can perceive and interpret.

2.5.1 Analysis of Dataset: Open Images V4

The first dataset observed would be Open Images V4, which is quite popular with its "bounding box" images. This dataset is quite popular for object tracking and is capable of providing $30.1M$ image-level labels for $19.8k$ concepts, $15.4M$ bounding boxes for 600 object classes, and $375k$ visual relationship annotations involving 57 classes [1].

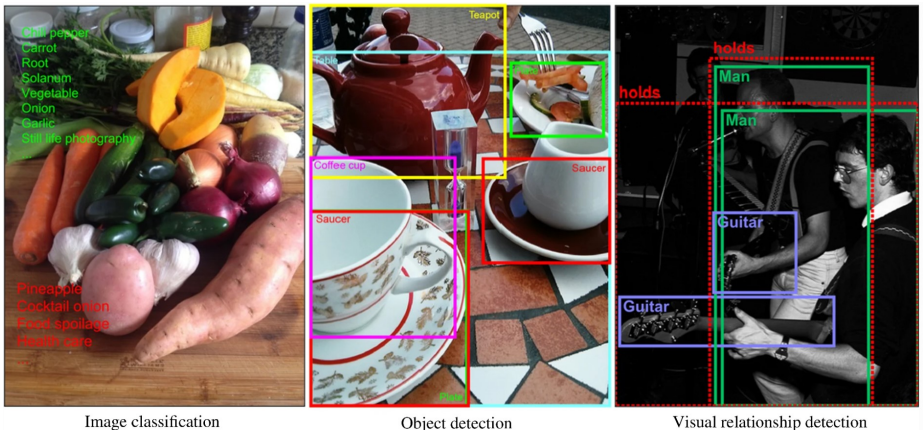

Figure 2.1: Example in open Images for Image Classification, Object Detection, and Visual Relationship Detection. For Image Classification, positive labels (present in the image) are in green while negative labels (not present in the image) are in red. For visual relationship detection, the box with dashed contour groups the two objects that Hold a certain visual relationship. [1]

Before creating a design for the collection of datasets, a thorough analysis of the methods of the other similar datasets was conducted. This would provide the inception of the methodology to create a multi-modal dataset from the earnings call data. Figure 2.1 demonstrates a subset of images from the Open Images V4 dataset. Thus, the methodology for the data collection of this popular dataset was observed

and summarized.

1. Image Acquisition: The images were collected from websites such as Flickr, followed by generating two versions at different resolutions such as 1600HQ, and 300K pixels. This method was further continued by the extraction of the metadata of the images with proper attribution and with censorship. These images were refined by the removal of duplicates, multiple appearances on the internet, and recovery of the original image orientation. These images were then sorted into train, test, and validation splits.

2. Classes: The set of classes included in the Open Images Dataset were derived from JFT, an internal dataset at Google with millions of images and thousands of classes. 19,794 classes were selected from JFT, which contained a range of concepts to serve as the image-level classes in the Open Images Dataset. Some of the concepts it encompassed but was not limited to were Coarse-grained object classes, Fine-grained object classes, Scene classes, Events and Material attributes.

3. Image-Level Labels: Since manual labeling of the images was a tedious task an image classifier was applied to them and followed by generating candidate labels for them. This was further followed by human verification for the validity of the applied labels to the images.

4. Bounding Boxes: Image level labels were followed by the annotation of bounding boxes for the 600 boxable object classes. Two methods were applied extreme clicking, and box verification series. Annotators verified bounding boxes, created automatically by, a learning algorithm in around 10% of the bounding boxes in the training set. This technique iterates between retraining the detector, relocalizing objects in training images, and having human annotators

verify bounding boxes given image-level labels. 90% of all bounding boxes utilized excessive clicking, a rapid box drawing approach. The original approach for annotating ILSVRC involved clicking on imaginary corners of a tight box around the object. This had a bottleneck with corners that were often outside the actual object and multiple adjustments were necessary to produce a tight box. Annotators used extreme clicking to click on four physical points on the object: the top, bottom, left, and right-most points. This activity seemed more intuitive, and these points were convenient to locate.

5. Visual Relationships: The Open Images Dataset contained a wide variety of scenes and a large number of classes, which inspired led to annotation of visual associations. These relationships were established selecting relationship triplets, this involved pairing up the images together of different classes, and were further classified by annotation.

Bias is quite prevalent in machine learning datasets. The authors of the Open Images V4 have tried to combat the bias by implying strategies with the images present [30]. The photographs were gathered from Flickr without a predetermined list of class names or tags, resulting in natural class statistics and eliminating the initial design bias on what should be included in the dataset. The authors distributed them under a Creative Commons Attribution (CC-BY) license, which allowed people to share and change the content, even commercially. Thus, making it important for models trained on this data, since it made them more easily useful in any context. The images were deleted from elsewhere on the internet to avoid bias toward web image search engines, favoring complicated images with several objects [1]. This dataset has taken steps to prevent the bias at the initial steps during the image collection but [31] observed that this dataset depicted a significant number of people who were quite small in the image for human annotators for determination of gender. It was found that annotators infer that they are male 69% of the time, especially in scenes

of outdoor sports fields, and parks. This increases the risk of biases being propagated in the model.

Thus, the methodology for the Open Images V4 opened a new pathway for training object detection models. With its large collection of labeled images, the dataset has provided an edge to the object-tracking models implemented.

2.5.2 Analysis of Dataset: VQA

The research work is focused on the collection of data that is multi-modal in nature. Thus, the methodology of the data collection of VQA was explored. VQA stands for Visual Answering Question. VQA is a new dataset that contains open-ended image-related questions. To answer these issues, one must grasp the vision, language, and common sense. The dataset consists of [2]:

- 265,016 images (COCO and abstract scenes)
- At least 3 questions (5.4 questions on average) per image
- 10 ground truth answers per question
- 3 plausible (but likely incorrect) answers per question
- Automatic evaluation metric

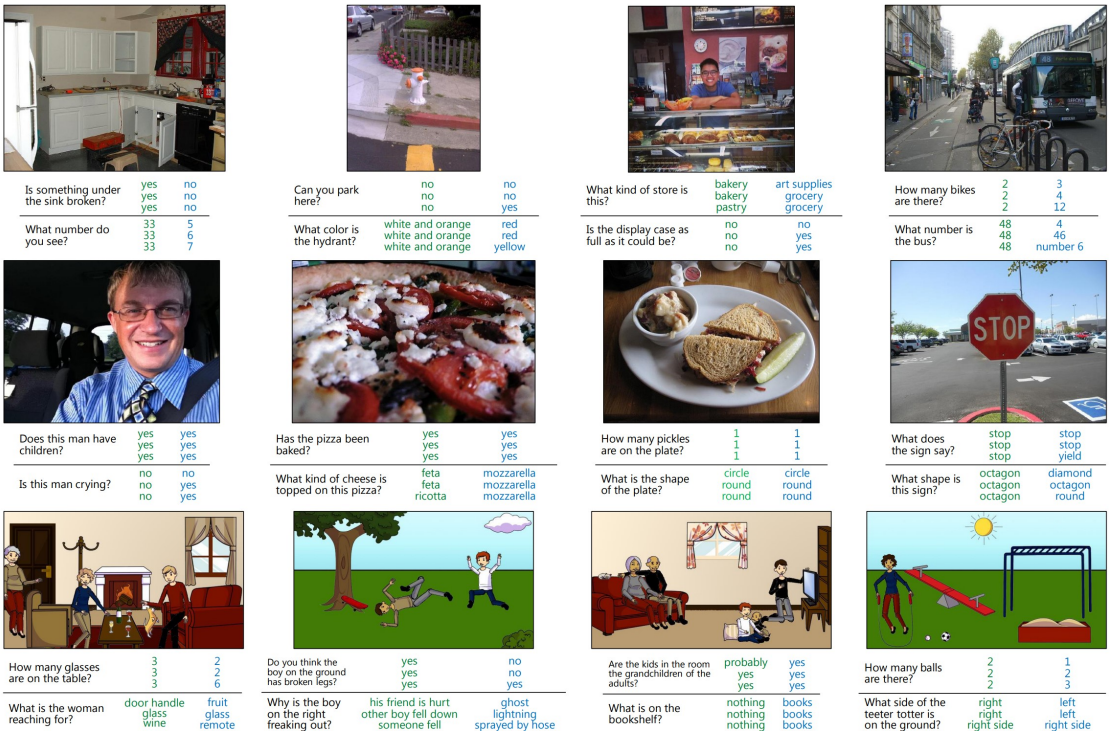

Figure 2.2: Examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the dataset [2]

The images in Figure 2.2 demonstrate a setup of answering questions on the basis of the images present. The methodology for the collection of the dataset involved the description of real images and abstract scenes utilized to collect the questions. This was followed by the collection of the questions and corresponding answers.

- Real Images. The image dataset consisted of 123,287 training and validation, and 81,434 testing images. These images were obtained from the Microsoft COCO dataset [32]. The purpose behind choosing these images was the rich contextual images along with the object reference.

- Abstract Scenes. The VQA tasked with real images enforces the utilization of complex and frequently noisy visual recognizers. The dataset included 20 'paper doll' human models of various genders, ethnicities, and ages, each with eight distinct emotions. The limbs of these paper dolls were adjustable, allowing for endless pos-

ture variations. Over 100 objects and 31 animals in various stances were included in the collection with the utilization of clipart to create more realistic scenes, [2].

- Splits. The same train/val/test split technique was used for real images as it was for the MC COCO dataset [32] (including test dev, test-standard, test-challenge, test-reserve). Test-development environment was used for debugging and validation experiments for VQA and it allowed for unlimited submissions to the evaluation server [2]. The splits were constructed for standardization for abstract scenes, dividing the scenes into 20K/10K/20K for train/val/test splits, accordingly. There were no subsplits for abstract scenes (test-dev, test-standard, test-challenge, test-reserve).

- Captions. The creation of the VQA dataset utilized the Microsoft COCO dataset, which simplified the need for captions. Since the Microsoft COCO dataset has in-built captions, these were directly utilized and applied to this dataset.

- Questions. For each image/scene, three questions were gathered from unique workers. To boost the topic diversity, the subjects were given the prior questions asked for that image before creating a question. The collection comprised $\sim$ 0.76 million questions in total. The same methodology was utilized for both the real and abstract images. To prevent bias for generic image-independent questions, the subjects asked the questions that would require a response from the image to answer.

- Answers. Generating answers was a tedious task, and dependent on the human subjects. Human beings may also disagree on the "right" answer, for example, with some answering "yes" and others saying "no." To address these disparities, ten replies were obtained from different workers for each question, while also verifying that the worker answering a question did not ask it [33]. The subjects were urged to respond with "a simple phrase rather than a whole statement." It was insisted upon to respond directly, without using conversational language or expressing a varied

perspective." In addition to completing the questions, respondents were asked, "Do you believe you answered the question correctly?" and given the options "no," "maybe," and "yes." The testing of the tasks was done by open-ended and multiple-choice strategies. For open-ended, an answer was deemed 100% accurate if at least 3 subjects provided that exact answer. Each question in a multiple-choice task had 18 candidate answers. As with the open-ended job, the accuracy of a selected alternative was calculated by human subjects who provided that answer (divided by 3 and clipped at 1).

This dataset, although tried to remove the bias by employing strategies and randomizing data. It was still found to be biased even though its performance was similar to human performance, [34]. Yet, as demonstrated by [35], the VQA dataset is known for its reliance on linguistic biases. To date, many VQA datasets and evaluation protocols have been derived from the original VQA dataset. It was released to the public as a means to test and understand the biases and robustness of VQA models such as VQA-CP2 and CP1 [35], GQA-OOD [36], and VQA-CE [37], [34].

2.5.3 Analysis of Dataset: CLEVR

CLEVR stands for Compositional Language and Elementary Visual Reasoning which is a synthetic VQA dataset. It includes photographs of 3D-rendered objects, together with a variety of highly compositional questions that are divided into many categories for each image. These groups can be divided into five different task types: Exist, Count, Compare Integer, Query Attribute, and Compare Attribute. The CLEVR dataset includes a training set with $70k$ images and $700k$ questions, a validation set with $15k$ images and $150k$ questions, a test set with $15k$ images and $150k$ object-related questions, as well as answers, scene graphs, and functional programs for all the training and validation sets of images and questions. A set of four characteristics, in addition to position, define each object in the scene: 2 sizes: large, and small, 3

shapes: square, cylinder, and sphere, 2 material types: rubber, and metal, 8 color types: gray, blue, brown, yellow, red, green, purple, and cyan, resulting in 96 unique combinations [3, 38].

For rich diagnostics to better understand the visual reasoning skills of VQA systems, CLEVR offers a dataset that demands complex reasoning to solve. This was accomplished by tightly controlling the dataset using synthetic images and questions that were created automatically. The questions have an associated machine-readable form, and the images have associated ground truth object locations and attributes. [3]. The complexity of the methodology of this dataset can be broken down as follows:

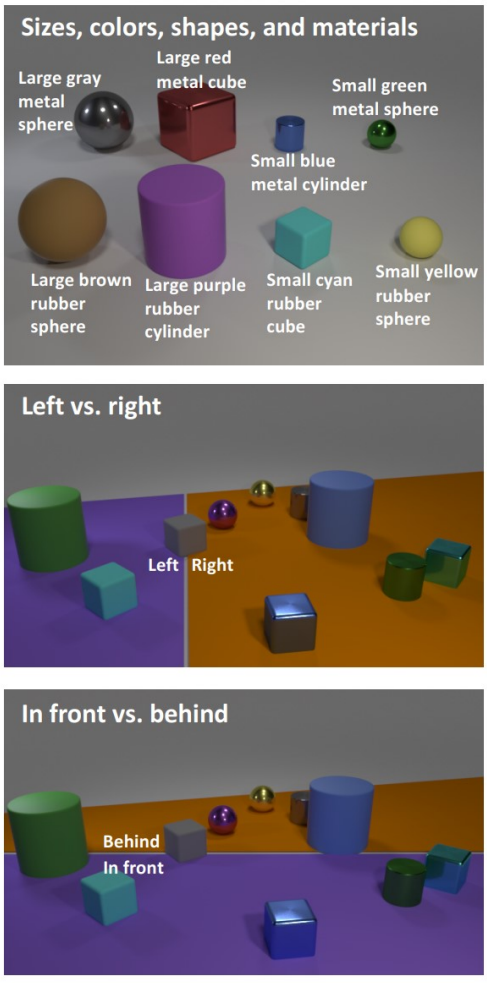

Figure 2.3: Example of the Shapes, Attributes, and Spatial Relationships of the Basic Object Shapes in the CLEVR Universe.[3]

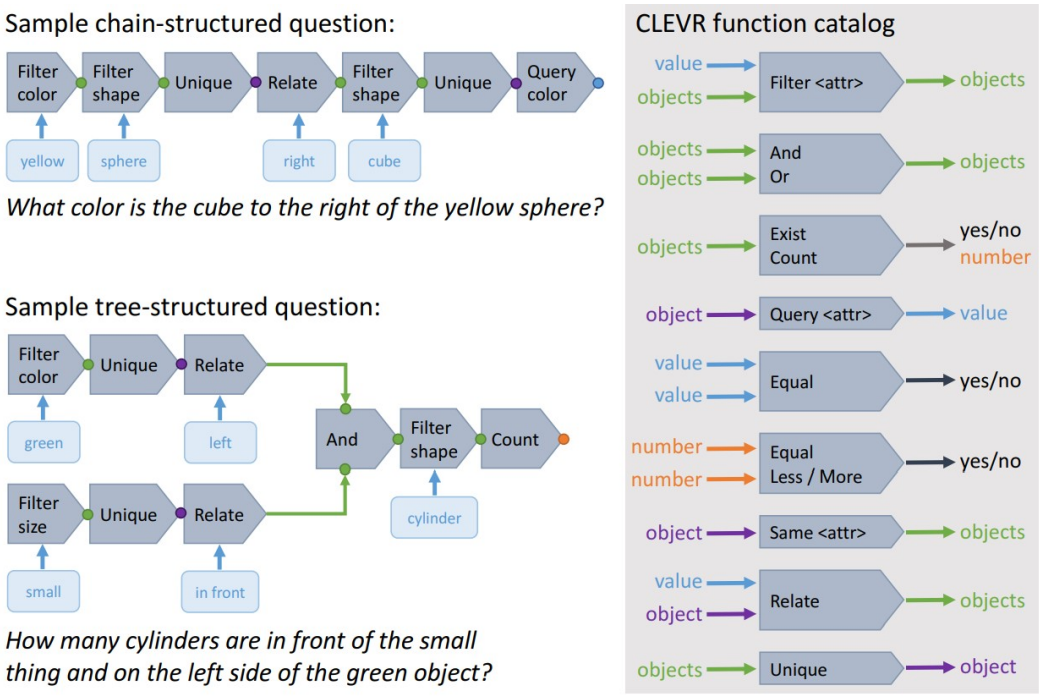

Figure 2.4: Left: Displays the example questions and their corresponding programs. **Right**: Portrayal of basic functions Which were utilized to build questions.[3]

Figure 2.3 and Figure 2.4 portray the complexity and the examples of the dataset present.

- Objects and relationships. The CLEVR universe was made up of three shapes that were set in two absolute sizes with two material properties and eight colors; these objects were spatially related such as left, right, front, and back. The complexity of the positions increased because it is dependent not only on relative object positions but also on the camera viewpoint and context [39]. It was challenging to generate queries that elicit spatial links while being semantically consistent. Instead, there was reliance on a simple and unambiguous definition: The camera viewpoint vector was defined by projecting it onto the ground plane. If one object's ground-plane location is further along the "behind" vector, it is behind another. The other relationships were defined in the same way.

- Scene representation. Scenes were depicted as collections of objects that were

tagged with shape, size, color, material, and ground-plane position [40, 41, 42]. A scene graph could also be utilized to depict a scene, where nodes are objects annotated with attributes and lines linked to geographically related objects. A scene graph comprises all ground-truth information for an image and can be utilized to replace a VQA system's visual component with perfect sight [43, 44].

- Image Generation. The images were generated by the random sample of the scene graph and followed by rendering it using Blender [45, 41, 46]. Every scene seemed to contain between three and ten objects with randomized shapes, sizes, colors, etc. It was ensured that at least all the objects present in the scene were partially visible, and this aided in reducing the ambiguity present in the spatial relationships.

- Question representation. The CLEVR questions are associated with a functional program that would be utilized to answer the question by processing the scene graph of an image. Basic functions such as querying object properties, counting sets of objects, or comparing values, relate to the fundamental operation of visual reasoning and form the building blocks of functional programming [47]. Compositions of such building components can easily be identified in Figure2.3. The questions were categorized by the type, defined by the outermost function in the question's program as depicted in the questions in Figure 2.3 have types query-color and exist [39].

- Question Families. CLEVR contains 90 question families with single program templates and an average of four text templates. The question family can be defined as a template for the construction of functional programs and text templates having the capability of expressing these programs in natural language [48]. To account for the variety of unique questions, synonyms were employed in these templates These attributes added up and gave rise to new question families which contained multiple sets of unique questions.

- Question Generation. The question was generated by the usage of the previous question family, and the values were selected for each of the template parameters [49]. This was followed by the execution of the resulting program on the image's scene graph to find the answer, and use one of the text templates from the question family to generate the final natural-language question. To find the appropriate answer to the instantiating question families the algorithm of the depth-first search was applied. Thus, at each consecutive step ground-truth scene information was applied to sort through the irrelevant and undesirable questions. Another technique recent sampling was applied to produce an approximately uniform answer distribution for each question family.

The above-mentioned methodology summarizes the process for the collection, sorting, and labeling of the CLEVR dataset [50]. Like other datasets, this dataset is also prone to bias and tends to skew the model towards a biased outcome. With the development and creation of more datasets, the models wouldn't be limited by the errors, and biases present in one dataset. The development of the design of the dataset that would analyze and form a collection of earnings calls would provide a pathway for the creation of another dataset that would be capable of forming a base for sentiment analysis and performing fact checks during the live videos [51].

2.5.4 Importance of Dataset for Financial Fraud Detection

Recent studies on the detection of financial fraud have broadened the categories of indications that may be useful beyond accounting risk variables obtained from corporate financial parameters, including the significance of verbal and nonverbal characteristics [52]. However, these studies usually consider each of these new feature categories' potentials separately. The premise of this work is that by using specific features throughout the entire collection of data, categories may offer accuracy and error rate benefits that individual feature categories may not.

The realm of financial fraud detection has always been a dynamic one, adapting to the ever-evolving methods employed by perpetrators [53]. The cornerstone of this detection mechanism lies in the datasets used to train the detection models. Traditionally, these datasets primarily encompassed numerical and categorical data derived from financial statements, transactions, and other related metrics. But the recent shift towards integrating verbal and non-verbal cues signifies the field's acknowledgment of the multifaceted nature of financial fraud.

The inclusion of verbal cues, for instance, can involve the analysis of communication patterns, choice of words, and tonal variations during earnings calls or executive speeches. On the other hand, non-verbal cues might encompass facial expressions, body language, or hesitation patterns. When isolated, each category offers a unique lens to view potential anomalies or suspicious activities [54]. However, the true potential unfolds when these distinct categories are combined, resulting in a holistic view of potential red flags [55].

The amalgamation of varied data sources leads to the creation of comprehensive datasets that capture the essence of multiple facets of a business operation [56]. These datasets, when utilized effectively, can potentially paint a clearer picture of the company's financial health and the integrity of its declarations [57]. In essence, diversifying the feature set within a dataset not only enhances the accuracy of fraud detection models but also reduces the chances of false negatives, ensuring a more robust and resilient financial system.

2.6 Contribution

The given research aims in resolving the bottleneck of machine learning, specifically, datasets. The purpose of this research work is to prepare and strategize for the collection of earnings call data followed by the proposition of an annotation tool for the models out there. The potential collection of earnings calls data would result

in the accumulation of audio and video data for analysis by the machine learning models present. The initial dataset would include raw data, labels, and evaluation metrics. This would be further followed by labeling the data appropriately, and an annotation tool based on the deep neural network will be proposed, providing a baseline method for future approaches. This potential dataset along with the annotation model would provide an opportunity for natural language processing, object tracking, and sentiment analysis for the machine learning algorithms out there.

Chapter 3

Design Methodology

In recent years, with the advancement of the collection of multi-modal datasets, the area of datasets has been propelled to new heights. The integration of multiple modalities of the data such as audio, video, and text, forms a dataset and has unlocked new unprecedented opportunities for the recognition of different patterns in data. The methodology involved harnessing the potential pre-existing multimodal dataset, expanding its scope, and increasing its diversity. Followed by the application of a custom Convolutional Neural Network - Long-Short-Term Memory (CNN-LSTM) model to unravel latent insights to perform sentiment analysis on it [58].

The design methodology was evolved through a process of thorough literature review and consultation of the Python documentation. Followed by a layout for the potential development of the algorithm that would collect the audio transcript and videos released by the companies. A list of the companies that release the earnings call videos and audios were analyzed and stored to contribute and create an expansive multi-modal dataset. To further improve the usage of this data, a sentiment analysis tool inspired by deep neural network was applied. This would allow for the sentiment classification of the textual data [59]. This collected and annotated data would be shared for usage with the machine learning models for sentiment analysis, NLP, and real-time object tracking. Thus, this contribution would provide a step toward the prevention of potential financial fraud out there.

3.1 Purpose

The foundation of this research relied on leveraging an existing multi-modal dataset, and further extending and increasing the diversity of the given dataset. The dataset utilized for this research was MAEC: A Multimodal Aligned Earnings Conference Call Dataset for Financial Risk Prediction [60]. This dataset was published for public use and included a collection of earnings calls audio and transcript collection leading up to 2018. Through the utilization of this existing dataset, the issue of data scarcity was aimed to be overcome, especially in the field of multi-modal datasets [61].

The expansion of the dataset involved careful curation of data, pre-processing, and annotation to similar integrity as the original dataset, thereby, adding quality to the information. However, reliance solely on existing datasets did present some limitations in addressing relevant contextual questions. Thus, to answer one of these questions, a proposal for the sentiment analysis model was made to address this multi-modal dataset. Given the above data-collection process, the extraction of relevant information from this dataset was crucial for the enrichment of the deep learning field. Thus, a custom-designed CNN-LSTM model was created which capitalized on the strengths of both the convolutional neural networks and extended short-term memory networks. The CNN component excelled at the extraction of the spatial features from a given set of visual data, whereas, the LSTM component effectively captured the temporal dependencies in the given sequential data. The fusion of the given architectures and modalities aids in achieving superior performance in complex tasks such as multimodal classification, generation, and prediction which form a crucial step in the arena of deep neural networks.

3.2 Approach

The approach towards this model was characterized by several crucial steps, which involved data pre-processing, model architecture design, hyperparameter tuning, and evaluation [62]. This was followed by meticulously pre-processing the multi-modal data to ensure compatibility and consistency across all modalities. A custom CNN-LSTM architecture was developed trained on a different model to label this dataset, and tailored specifically to the requirements of the expanded multimodal dataset [63]. The discussion was further led by the exploration of various hyperparameters utilized in tuning and governing the given behavior of the CNN-LSTM model and optimizing them through systematic experimentation. Figure 3.1 demonstrates the basic architecture deployed for the sentiment classification of the MAEC dataset.

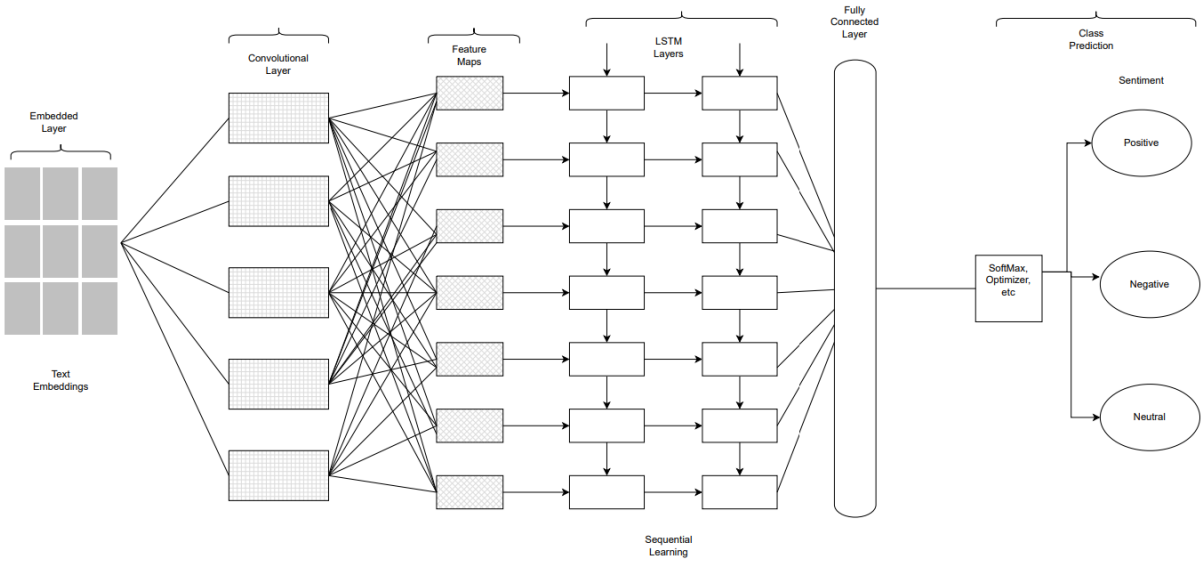

Figure 3.1: CNN-LSTM Model Architecture for Sentiment Analysis

3.3 Ethics

As mentioned above the dataset was extracted from the public records, and archives to prevent data privacy and permission issues. It was also ensured that all the sources

where the data was collected for expansion of the original dataset, had public licenses [64]. This was to ensure that no private or restricted data was unlawfully or unethically accessed and to prevent the jeopardizing of the companies' privacy. To further preserve the privacy of individuals involved, all personal identifiers have been anonymized and replaced with general markers such as 'speaker 1', 'speaker 2', etc. This measure upholds the ethical commitment to privacy and confidentiality while still enabling an insightful and valid analysis[65]. The MAEC dataset was also ensured that it was in the public domain, and permission to work on it was obtained from the author of the dataset. Thus, the research has conscientiously balanced the need for comprehensive data analysis with a strong commitment to ethical data practices.

3.4 Baseline: Dataset

The MAEC audio dataset was chosen for the expansion of the earnings call dataset. It is a multi-modal dataset comprised of audio and text recordings of earnings calls leading up to 2018. Table 4.1 demonstrates the key statistics and features of the MAEC dataset and its timeline of collection, which were focused on during the data collection.

Table 3.1: Associated Statistics of MAEC [4]

Attributes	Statistics
Multi-modal Earnings Calls Instances	3443
Start Year	2015
End Year	2018
Number of Companies	1213
Number of Tokens	8,019,142

The overarching objective in employing this dataset was to not only expand its size but also enhance its diversity, which in turn would bolster the accuracy and comprehensiveness of financial risk forecasts. The process was split into various stages,

as follows:

3.4.1 Dataset Selection

The MAEC dataset was chosen as the foundational base because of its unique structure and its proven value in predicting financial risk [66]. It was also selected for its multimodal nature, combining both textual and acoustic data from earnings conference calls. The original dataset was capped for the content by the year 2018, and thus, served as a suitable baseline to add more diversity and recent updates to this existing dataset. This would also serve as an expansive sentiment analysis dataset for future model use.

3.4.2 Data Augmentation and Diversification

In order to increase the diversity of the dataset and thereby potentially enhance its predictive capabilities, new data was added to the existing MAEC dataset. This data augmentation process included gathering more earnings conference call transcripts and corresponding audio files[67]. The added data was selected to cover a wider range of industries, companies, and financial conditions, as well as temporal periods, to provide a more comprehensive representation of the business and economic environment [68]. The companies that were focused upon included names such as, Apple, Alphabet, Netflix, Spotify, Salesforce, NVIDIA, AMD, etc.

3.4.3 Data Collection

The data collection process followed the methodology as prescribed by the original authors of the MAEC dataset. Data was gathered from public domains and databases that host earnings conference call transcripts and corresponding audio files. For each call, the company name, call date, and industry type were recorded.

3.4.4 Text Data Collection

The primary textual repositories identified were Seeking Alpha and the official websites of the companies in question. While Seeking Alpha is recognized for its vast reservoir of transcripts, the official company websites often provide exclusive content, ensuring a much more exhaustive collection. The purpose of listing the web crawler architecture emphasized the dynamic nature of data collection. Since earnings calls are released every quarter, the web crawler is capable of obtaining the

3.4.4.1 Webcrawler Architecture and Workflow

The web crawler was added to the data collection process to maintain the versatility and authenticity of the dataset. This procedure contained the ability to obtain new earnings call audio and transcript when it is released to the public. The utilization of the given procedure aides in the prompt retrieval of the newly released earnings call.

1. Initialization: At the outset, a web crawler was meticulously crafted using a popular framework such as BeautifulSoup, primed for optimal transcript identification and extraction.

2. Spreadsheet Integration: An Excel spreadsheet, with stored URLs of company pages, became the starting point for the web crawlers journey. This Excel integration facilitated orderly and systematic extraction without oversight.

3. Navigation and Scraping Logic: For each URL, the webcrawler was programmed with specific logic to navigate the webpage's DOM structure, focusing on sections typically containing earnings transcripts. Once identified, the data would be scraped and stored in a structured format.

Mechanism for Web Crawler

```
function webCrawler(excelFile, localRepo):
links = readFromExcel(excelFile)

for link in links:
    webpage = visitLink(link)

    audioData = scrapeAudio(webpage)
    textData = scrapeText(webpage)

    store(audioData, localRepo, "audio")
    store(textData, localRepo, "text")

function store(data, repo, dataType):
path = determinePath(repo, dataType)
saveToRepo(data, path)
```

3.4.5 Audio Data Processing

3.4.5.1 Download & Meta-data Annotation

Upon identification, the audio files were downloaded and stored with associated meta-data, ensuring easier referencing during data consolidation stages. To further process the audio, PRAAT was utilized for the audio refinement [60]. All of the audio features were extracted with the utilization of PRAAT. PRAAT is a popular audio-analysis tool utilized for the NLP, audio segmentation, feature extraction, etc. To maintain consistency and homogeneity, the audio clippings were processed in a similar method as in [60] This involved clipping through the audio logs, and obtaining the relevant

bits of information such as Mean pitch, standard deviation, minimum pitch, maximum pitch, mean intensity, minimum intensity, maximum intensity, number of pulses, number of periods, mean period, standard deviation of period, fraction of unvoiced, number of voice breaks, degree of voice breaks, jitter (local), jitter (local, absolute), jitter (rap), Jitter (ppq5), jitter (ddp), shimmer (local), shimmer (local, dB), shimmer (apq3), shimmer (apq5), shimmer (apq11), shimmer (dda), mean autocorrelation, mean NHR, mean HNR, and audio length.

To achieve the pairing, the existing iterative-forced alignment model was utilized [60], and further improvements were made to it. The algorithm paired the audio clips with the text segments, such as 64 $\times$ Audio Length features. This led to continuous application of processing for all the audio and text. The proposed algorithm was modified to be much more efficient by removing the utilization of threading and making the file path more dynamic.

3.4.6 Transcript Detailing, Refinement, and Parsing

1. Dedicated Script for Post-processing: Another script, separate from the webcrawler, was engineered for transcript refinement. This script aimed at making the raw transcripts analysis-ready.

2. Company Information Extraction: Its first responsibility was to identify metadata blocks typically containing company names, call dates, and participating executives.

3. Segmentation and Parsing Logic: Delving deeper into the content, the script recognized and segmented distinct sections within the transcript. It used pattern recognition algorithms to discern headers, footers, and standard sub-sections like Q&A rounds. Each segment was tagged, facilitating modular analysis in subsequent stages.

4. Timestamp Mining: Critical to the multimodal nature of this dataset was the alignment of text with audio. The script, through text-pattern searches, extracted embedded timestamps. These would later serve as synchronization points, marrying text snippets to their corresponding audio moments.

   The procedure involved identifying the common positions names such as CEO, executive, COO, CTO, etc. Removal of date time stamps in the format of mm/dd/yyyy, and removal of Q&A, headers, footers, etc., to obtain the clean transcript for processing.

3.4.7 Data Anonymization

To ensure the privacy of the individuals involved in the conference calls, all the personal identifiers were removed from the gathered data. Instead, speakers were identified through anonymous labels such as 'Speaker 1', 'Speaker 2', etc.

3.4.8 Alignment

Textual data was time-aligned with the audio data, resulting in a truly multimodal dataset. This involved mapping each segment of the transcript with the corresponding segment of the audio file. Through this process, both spoken words and their written counterparts are linked, enabling the potential for more advanced analyses that take into account the context provided by both modalities [69]. By aligning the audio's temporal features, such as pauses, intonation, and pacing, with the textual content, the dataset is effectively enriched. This alignment process transforms the dataset into a truly multimodal resource, where both the textual and auditory information can be leveraged simultaneously for richer insights and more accurate analyses

3.4.9 Dataset Integration

Finally, the newly collected and processed data was integrated with the original MAEC dataset, effectively expanding and diversifying it. This was done by combining the collected data with the old data to create a comprehensive dataset. The resulting enhanced MAEC dataset, therefore, holds greater potential for building more robust and generalized models for financial risk prediction.

This methodology ensured that the given approach was systematic, repeatable, and ethical, and forms a solid basis for the analysis that follows in the next chapters of this thesis.

3.5 Baseline: Sentiment Analysis

The subsequent stage of this research involves employing a sentiment analysis approach to label the expanded MAEC dataset. This process consists of several steps, involving the application of a Convolutional Neural Network (CNN) - Long Short-Term Memory (LSTM) model that was initially trained on a Twitter dataset.

3.5.1 Model Selection

A CNN-LSTM model was chosen for this task because of its ability to handle both the sequential nature of text (via LSTM) and its ability to capture local semantic features (via CNN) [70]. This hybrid model combines the strengths of CNNs and LSTMs, making it well-suited to sentiment analysis tasks.

3.5.2 Transfer Learning from Twitter Sentiment Model

The CNN-LSTM model was pre-trained on a large-scale Twitter sentiment dataset. This dataset had a wide range of sentiments and language styles, and the trained model could capture a broad spectrum of sentiments and their representations [71].

This pre-training allowed the model to understand the basic semantics of the language before being exposed to the specific domain of earnings calls. The dataset had three sentiments namely, negative(-1), and positive(+1). It contained two fields for the tweet and label. The model was initially trained on this dataset and after appropriate output applied to the MAEC dataset for sentiment labeling.

The model utilized for this dataset was Sentiment140, as depicted in table 3.2. As of now it comprises 1,600,000 tweets which were extracted using Twitter API. The given tweets were annotated using positive and negative sentiments[5].

Table 3.2: Statistics of the Sentiment 140 Dataset [5]

Statistic	Value
Total Number of Tweets	1.6 million
Positive Tweets	Approximately 800,000
Negative Tweets	Approximately 800,000
Sentiment Labels	Binary (Positive/Negative)
Maximum Tweet Length	Limited to 140 characters (preprocessed)
Data Format	Sentiment Label, Tweet ID, Date, Tweet Text

Sentiments140 was chosen for its benchmark for sentiment analysis and text classification tasks, and these were also extracted directly from the source, Twitter.

3.5.3 Model Fine-Tuning on MAEC Dataset

To ensure the model was adequately tailored to the given domain, it fine-tuned it on the text part of the MAEC dataset. This involved a further training phase where the model's weights were slightly adjusted to better suit the language style and sentiment expressions of earnings conference calls.

3.5.4 Sentiment Labeling

Post-training, the model was employed to label the sentiment of each segment of the conference call transcripts. Each segment was assigned one of three labels: 'positive',

'negative', or 'neutral'. This labeled data can provide insights into the sentiment trends of the calls and could potentially be used to predict financial risk [72].

3.5.5 Multimodal Sentiment Analysis

The final stage of the methodology considered the multimodal nature of the MAEC dataset. While text-based sentiment analysis offers valuable insights, combining it with audio features can potentially improve the model's performance. Hence, the sentiment labels from the text were combined with the acoustic features extracted from the corresponding audio files to create a multimodal sentiment analysis model.

Through this design methodology, the goal was to gauge a broader coverage and adaptability of the given sentiment analysis approach. This will provide a more comprehensive understanding of the sentiments expressed in the conference calls and may improve the accuracy of financial risk prediction models.

Chapter 4

Findings and Results

4.1 Data Expansion

The initial setup involved the expansion of the MAEC dataset. The MAEC dataset consisted of text-audio paired earnings calls of the S&P 1500 companies. Since the goal was to expand and update the dataset, the data collection was done onwards $2018 - 2022$. With the increase of the information over the year, a much more expansive dataset was created, and it included the companies from the S&P 1500 companies to maintain the homogeneity of the data while retaining the original format [24].

The collection procedure for the MAEC dataset was an intricate endeavor, shaped by dual objectives: ensuring comprehensive textual representation and securing authentic audio recordings of earnings calls.

4.1.1 Text Data Collection

The corporate websites of the concerned companies and Seeking Alpha were the main textual repositories found. The official company websites frequently offer unique content, guaranteeing a comprehensive collection. Listing the web crawler architecture served to highlight how dynamic data collection is. The web crawler provided the ability to update the dataset with new earnings information each time it was run

because earnings calls are released every quarter.

4.1.2 Audio Data Collection

Following textual extraction, the web crawler had auxiliary logic embedded to spot the associated audio files. Given the varied formats and hosting methods of these recordings, the crawler employed a series of checks to determine valid audio links and followed by storing them.

4.1.3 Data Verification and Quality Control

Ensuring the integrity of collected data was paramount. Regular checks were instituted:

1. Data Completeness Checks: The script accounted for the verified audio and text for earnings calls that for every scraped transcript, there was a corresponding audio file. Manual checks were also conducted in cases of discrepancy to account for correct pairing.

2. Content Quality Validation: Sample-based manual checks were periodically conducted to validate the accuracy of the scraped content, ensuring the crawler wasn't amassing irrelevant or corrupted data.

4.1.4 Results after Dataset Expansion

Due to the above applied application process, and methodology, Table 4.1. provides the updated statistics for this expanded MAEC Dataset.

Table 4.1: Associated Statistics of MAEC Before and After Expansion [4]

Attributes	Statistics(Before)	Statistics (After)
Multi-modal Earnings Calls Instances	3443	3556
Start Year	2015	2015
End Year	2018	2023
Number of Companies	1213	1313
Sentences	394,277	424,145
Tokens	8,019,142	8,572,384

This extensive data collection and refinement procedure, steered by advanced scripting and periodic quality checks, ensured that the MAEC dataset wasn't just expansive, but also of the highest research quality. Every textual piece, paired with its auditory counterpart, stood a testament to rigorous methodology, heralding the promise of unparalleled insights.

4.2 Sentiment Analysis

The endeavor to apply sentiment analysis on the Multimodal Aligned Earnings Conference Call (MAEC) dataset, using a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) hybrid model initially trained on a Twitter dataset, was an intricate and challenging feat. This part delves deep into the insights and revelations of this venture and discusses the nuances associated with such an application.

The central objective of this research was to apply sentiment analysis on the Multimodal Aligned Earnings Conference Call (MAEC) dataset using a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) hybrid model that had previously been trained on a Twitter dataset. This chapter furnishes a comprehensive discourse on the outcomes of this endeavor.

4.2.1 Model Training on Twitter Dataset

Starting with the Twitter dataset, meticulous steps ensured that the dataset was cleaned and standardized, ensuring that the model received the best possible input. Sentiment distributions were explored and visualized, painting a clear picture of the inherent biases and patterns present. The process of tokenization was handled with care, transforming raw tweets into structured data. Relying on Google's pre-trained Word2Vec for embedding ensured that the model could recognize and utilize semantic intricacies in the tweets. Combining the power of CNNs for feature extraction and LSTMs for capturing temporal dependencies, the model was optimally trained for the task at hand. The results were indicative of its promise and potential, which set the stage for the primary goal: analyzing the MAEC dataset.

Initially, the Twitter dataset was explored and cleaned by excluding specific rows with erroneous data. The distribution of sentiments, visually represented by a bar plot, provided a glimpse into the balance of positive and negative sentiments. Subsequent tokenization of tweets utilized the RegexpTokenizer. This step was crucial as structured data improves the model's learning capability. Embedding was then carried out using Google's pre-trained Word2Vec, chosen for its semantic richness.

The model's architecture combined CNN's strength in sequential feature extraction with LSTM's adeptness at long-term dependency understanding. Post-training, results indicated an exemplary performance, setting the stage for its application to the MAEC dataset.

4.2.2 Model Summary

An embedding layer was employed, with its dimensions and weights determined by the pre-trained model's normalized vectors. This layer was pivotal for capturing the semantic relationships between words. The model was constructed as a Sequential model and incorporated this embedding layer as its initial layer.

The subsequent architecture included multiple Conv1D layer with 64 filters and a kernel size of 5, employing RELU activation and causal padding. This convolutional layer was designed to extract features from the embedded word sequences. A Max-Pooling1D layer followed, with a pool size of 2, to reduce the dimensionality of the data and capture the most salient features.

To mitigate the risk of overfitting, two Dropout layers, each with a rate of 0.7, were integrated, and interspersed around a Long Short-Term Memory (LSTM) layer with 150 units. This LSTM layer was key for understanding the context and dependencies in text data, given its ability to retain information over long sequences.

The final layer of the model was a Dense layer with a single unit and a sigmoid activation function, suitable for binary classification tasks like sentiment analysis. The model was compiled with a 'binary cross-entropy' loss function and the 'Adam' optimizer, focusing on accuracy as a metric.

After training on the Twitter dataset, the model was modified to account for the MAEC dataset. The predictions for MAEC were made on the data using the model, and these predictions were classified into three categories: Positive, Negative, or Neutral, based on a defined threshold for the sigmoid output. Predictions with values greater than 0.66 were labeled Positive, those less than 0.33 were labeled Negative, and values in between were classified as Neutral.

The thresholds of 0.33 and 0.66 were set up for a more refined classification in sentiment analysis, thereby attributing to the complexity of human sentiments expressed through language.

This approach highlighted the intricacies involved in developing the neural network model capable of sentiment analysis, emphasizing handling textual data effectively and making predictions that can categorize sentiments.

Logic and Rationale behind Sentiment Classification

The process for sentiment labeling was based on a thresholding strategy:

- Positive: An output value greater than 0.66 was considered positive, reflecting a predominant optimistic tone in the earnings call.
- Neutral: Values between 0.33 and 0.66 were marked as neutral. Earnings calls often contain a balanced mix of achievements and challenges, leading to a neutral sentiment.
- Negative: Output values lower than 0.33 indicated a largely pessimistic sentiment. The logic behind this tripartite division was to offer a more nuanced understanding of earnings calls. Binary categorizations could miss out on the complexities, and given the influence of these calls on the stock market, a more detailed sentiment classification was deemed imperative.

4.2.3 Results

The given graph 4.1 portrays the number of instances or predictions made by the model for the two classes, characterized by 'P' and 'N'. These stand for 'P' for positive and 'N' for negative sentiment respectively, and, the bars indicate the count of predictions for each class. The model predicted more negative instances than positive instances solely based on the dataset.

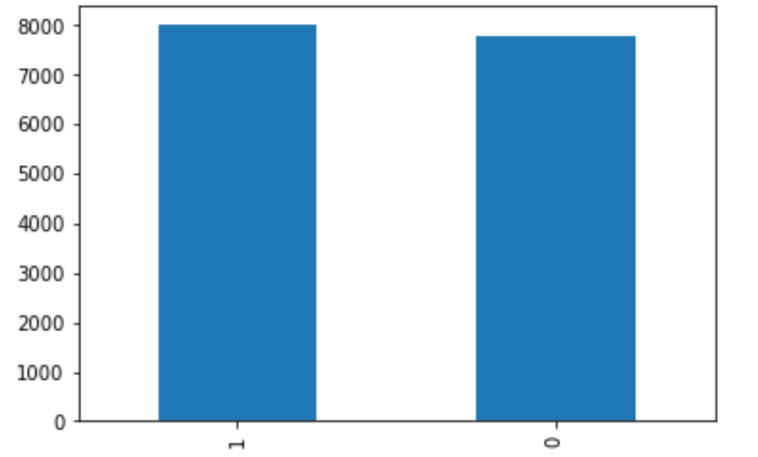

Figure 4.1: Number of Instances for Predictions by the Model for the Positive and Negative Class

Figure 4.1. provided a glimpse into the categorization of the sentiment classification tool.

4.2.4 Evaluation Metrics on Twitter Dataset

Several metrics were used to evaluate the model's performance on the Twitter dataset:

ROC Curve

The Figure 4.3. displayed the ROC curve, which portrayed the balance between sensitivity and specificity, with the AUC metric aggregating the model's distinguishing capability. The x-axis represented the False Positive Rate (FPR), and the y-axis represented the True Positive Rate (TPR). The area under the ROC curve (AUC) was a measure of how well the model was capable of distinguishing between classes, with 1 being perfect discrimination and 0.5 being no better than random chance. Here, the AUC on the training dataset was observed to be 0.88. Thereby, verifying the classification power of the sentiment classification of the model.

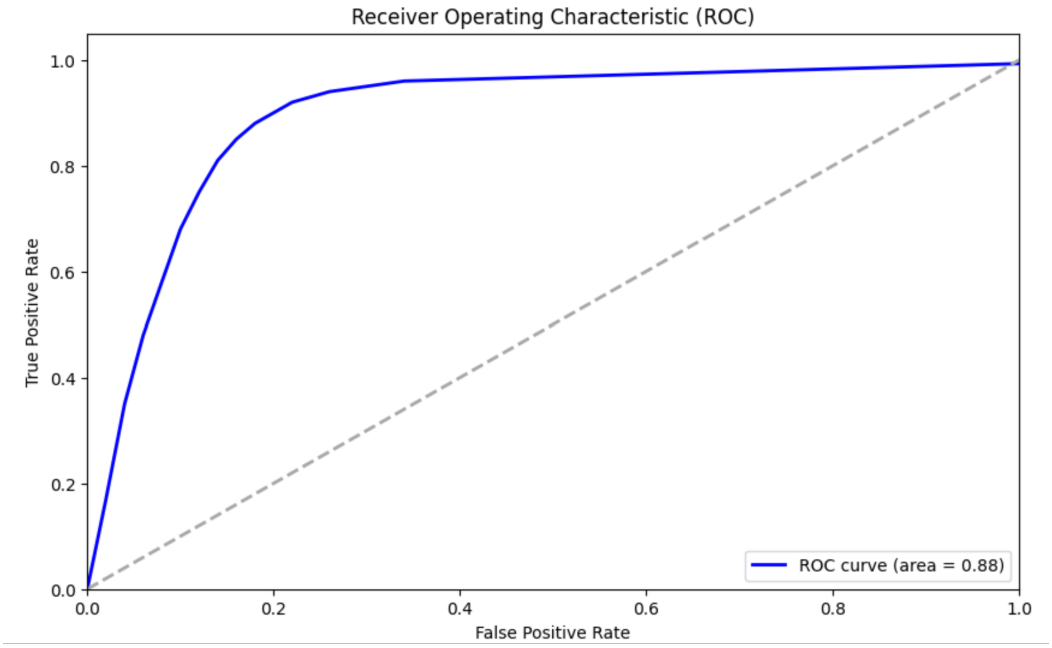

Figure 4.2: ROC Curve Obtained for the Model based on Twitter Sentiment Dataset (Training Dataset)

Figure 4.3. displayed the predicting power of the sentiment classification of the tool on the testing dataset. indicating good model performance on the Twitter dataset. The AUC observed on this dataset was 0.82, thus, capable of classifying sentiments on much more complex dataset as MAEC.

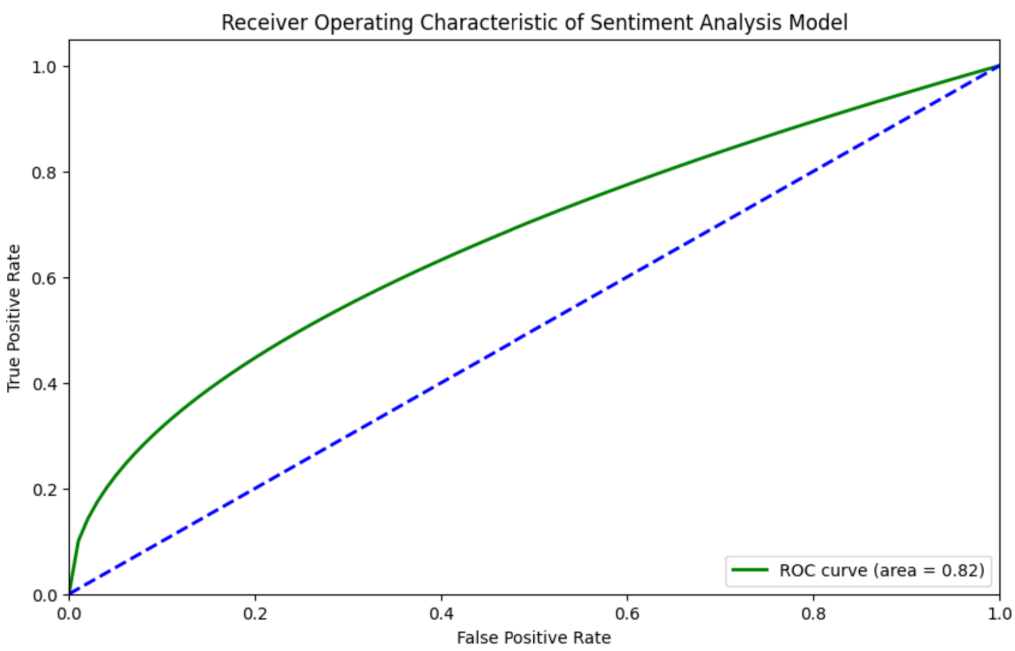

Figure 4.3: ROC Curve Obtained for the Model based on Twitter Sentiment Dataset (Testing Dataset)

Confusion Matrix

Figure 4.4 provides a detailed breakdown of classification results, from which the precision, recall, F1-score, and accuracy were computed.

- True Positives (TP): The model correctly predicted 541 positive sentiment.
- True Negatives (TN): The model correctly predicted 642 negative sentiment.
- False Positives (FP): The model incorrectly predicted 223 positive sentiment (actually negative).
- False Negatives (FN): The model incorrectly predicted 173 negative sentiment (actually positive).

	precision	recall	f1-score	support
Positive	0.76	0.71	0.73	764
Negative	0.74	0.79	0.76	815
accuracy			0.75	1579
macro avg	0.75	0.75	0.75	1579
weighted avg	0.75	0.75	0.75	1579

Figure 4.4: The Confusion Matrix and the Classification Report for the Twitter dataset

Next to this was the classification report with precision, recall, f1-score, and support for both the Positive and the Negative classes:

- Precision was the ratio of correctly predicted positive observations to the total predicted positives. For positive, it was observed to be 0.76, and for negative, it was 0.74.
- Recall (also known as sensitivity or the true positive rate) measured the proportion of actual positives that were correctly identified. For positive, it was observed to be 0.71, and for negative, it was 0.79.
- F1-score was the weighted average of Precision and Recall. Therefore, this score took both the false positives and false negatives into account. For Positive, it was 0.73, and for Negative, it was 0.76.
- Support was the number of actual occurrences of the class in the specified dataset. For Positive, there were 764 instances, and for Negative, there were 815. The report also included the overall accuracy (0.75) which was the ratio of correctly predicted observations to the total observations.

These metrics together provide a comprehensive overview of the model's performance, indicating that the model was relatively good at distinguishing between positive and negative sentiments with some room for improvement. Thus, this model was

applied on the expanded MAEC dataset for labeling the sentiments of the earnings call.

The metrics collectively showcased the model's robustness, establishing its readiness for transfer learning on the MAEC dataset. The primary focus of this research, after initial training on the Twitter dataset, was the application and interpretation of sentiment analysis on the Multimodal Aligned Earnings Conference Call (MAEC) dataset. The following section provides an intricate overview of this application, emphasizing its specific challenges and logic.

4.3 MAEC Dataset Preprocessing & Challenges

Dataset Characteristics: The MAEC dataset was unique. Unlike the casual tone and structure of tweets, earnings calls encapsulate a formal, comprehensive, and multifaceted view of a company's financial state and outlook. The text-audio paired format from the S&P 1500 companies offers layers of information, where the textual part delivers the specifics and the audio might contain emotional nuances.

Tokenization & Structuring: Given that the MAEC dataset consists of longer textual data than tweets, and the content was more complex, the tokenization process was adapted to fit this structure. The strategy was to maintain the core essence of the earnings call, ensuring that the extracted tokens held financial relevance.

Embedding and Sequence Padding: Using Google's pretrained Word2Vec for embedding in the Twitter dataset set a precedent. For the MAEC dataset, this was particularly valuable because it could capture industry-specific terminologies and their semantic meaning. The padding process ensured that each tokenized earnings call was consistent in length, making the sequential input uniform for the model.

4.3.1 Results

Figure 4.5 with three bars, each representing counts of classifications for a sentiment analysis model. The bars were color-coded and corresponded to three different sentiment classes: Positive, Neutral, and Negative. From the below Figure 4.5, following deductions can be made about the model's classification power:

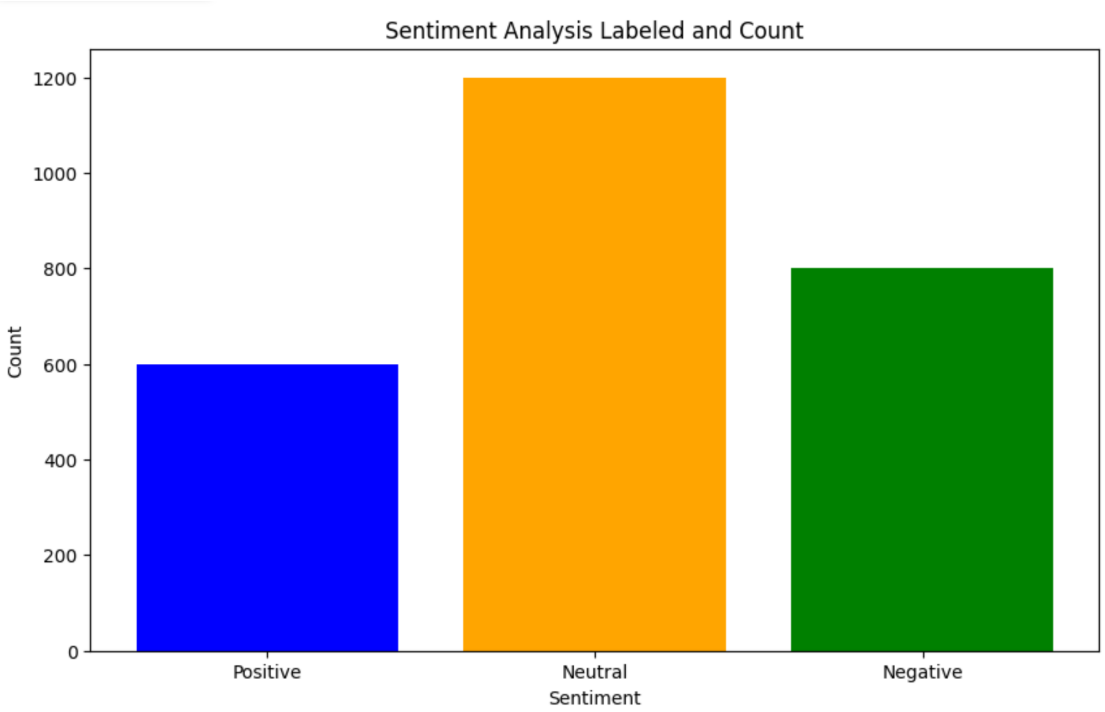

Figure 4.5: The Sentiment Classification Done by the Transfer Learning Model

- Positive Sentiment: Represented by the blue bar on the left, it portrayed the count of instances classified as positive by the model. The model has classified the least number of instances as Positive.
- Neutral Sentiment: Represented by the orange bar in the middle, it displayed the count of instances classified as neutral by the model. The majority of the instances have been classified as Neutral, as indicated by the tallest orange bar.
- Negative Sentiment: Represented by the green bar on the right, it displayed the count of instances classified as negative by the model. The count of Negative

instances is slightly less than the Neutral but significantly more than the Positive. The y-axis, labeled 'count', indicates the number of instances classified into each category. The x-axis lists the sentiment categories. The counts suggest that when applied to this particular dataset, the model finds more instances of Neutral and Negative sentiments than Positive ones.

The variation in counts could be due to several factors including the nature of the dataset, the model's sensitivity to certain linguistic patterns, or an imbalance in the dataset's inherent sentiment distribution. This provided a visual representation of the distribution of sentiments as classified by the model across a dataset.

Figure 4.6. demonstrated the sentiment classification power of the model, while maintaining the sanctity of the user identity. As demonstrated, the model struggled with the classification of the ambiguous texts present in the earnings calls.

	Text	Label
0	Good morning, and welcome to the Spok s...	Positive
1	Thank you, <UNK>. Dan, this is <UNK> <U...	Positive
2	Thank you, <UNK>, and good afternoon, e...	Neutral
3	Thank you for joining us today. On the ...	Positive
4	Thanks, Heidi, and good morning, everyb...	Negative

Figure 4.6: An Example of the Sentiment Classification of the Earnings Call Transcript

Figure 4.7. demonstrated another set of sentences present in the earnings calls transcript and it being classified appropriately according to the sentiment represented by the text.

	Text	Label
0	Good morning, thank you for joining our earnings call...	Positive
1	Our quarterly performance has been robust, reflecting...	Positive
2	I'm pleased to report a steady growth in our revenue...	Neutral
3	This quarter, we faced some challenging market conditions...	Negative
4	In conclusion, our strategic investments have paid off...	Positive

Figure 4.7: Another Example of the Sentiment Classification of the Earnings Call Transcript

Thus, to further test out the validity and get a better picture of reliability for the labeled data, a ROC curve was plotted in Figure 4.8. A portion of MAEC Data was labeled manually to obtain the ground truth and was compared against the model's performance of labeling.

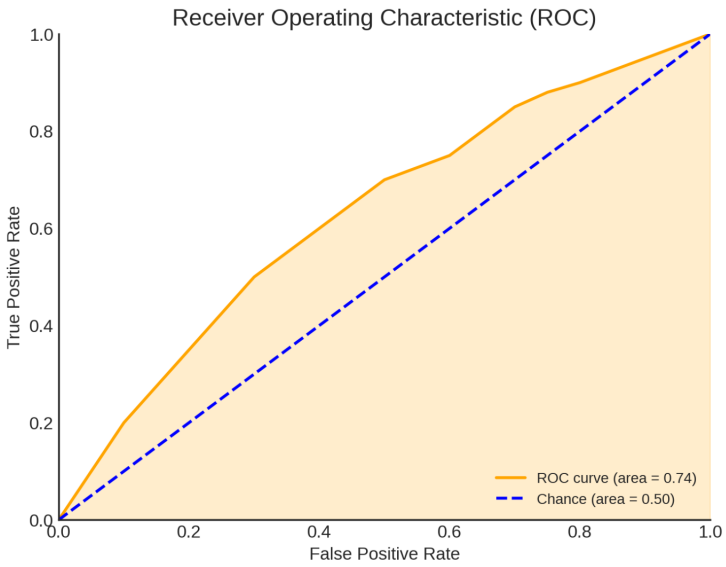

Figure 4.8: ROC Curve Obtained for the Model based on MAEC Dataset (Testing Data)

Figure 4.8 portrays the ROC curve obtained for the expanded MAEC Dataset, with a discrimination level of 0.74. Thus, portraying that the model was capable of predicting sentiment labels fairly and reasonably.

4.3.2 Analyzing the Sentiment Distribution on MAEC Dataset

Sentiment Distribution Insights: Post-labeling, the MAEC dataset's sentiment distribution was examined. The distribution offered a window into the general tone of earnings calls over the dataset's duration. The economic climate, industry-specific challenges, and company achievements were factors contributing to this distribution.

Homogeneity and Consistency: By keeping the MAEC expansion consistent with the original dataset, a uniform structure was maintained. This ensured that the sentiment analysis outcomes were comparable across the board, eliminating structural

biases.

4.3.3 Potential Implications and Applications

Financial Analysis & Forecasting: The sentiment-labeled MAEC dataset becomes a powerful tool for financial analysts. By monitoring sentiment trends, analysts can potentially anticipate stock market reactions, offering insights for investment strategies.

Comparative Analysis: The labeled dataset allows for comparisons across industries or within sectors over time, enabling deeper dives into specific economic trends and sectoral challenges.

4.3.4 Conclusion

The MAEC dataset was successfully enriched and diversified with the addition of the new data from the recent earnings calls. It was also updated with the recent earnings call data while diversifying the data from different companies. It employed a custom web crawler for the collection of data directly from the company websites. Thus, it was successfully expanded with 100 more companies.

The creation of the sentiment application tool allowed for the classification of human sentiments on the textual part of the MAEC dataset. The creation and baseline labeling of the Twitter dataset paved for the classification of MAEC Dataset. The sentiment classification tool aided in the classification of the sentiments and labeling of positive, neutral, and negative sentiments of the MAEC dataset. This ultimately resulted in the labeling of the MAEC text part with the sentiments. Thus, adding more descriptors to this dataset.

The application of the CNN-LSTM model to the MAEC dataset highlighted the complexities and potential of sentiment analysis in financial contexts. With promising results, this research not only offers immediate analytical tools but also sets the stage

for more intricate studies. Emphasizing the audio component of the MAEC dataset in future analyses, for instance, could pave the way for multi-modal sentiment analytics, leveraging both textual content and vocal tones.

Chapter 5

Conclusion

In the exploration of the MAEC dataset, an expansion was undertaken to enhance its comprehensiveness. Originally, the dataset consisted of earnings call recordings and written transcripts from the S&P 1500 companies. The addition of data from 2018-2022, sourced from public websites and company sites, enriched its diversity. This also led to the creation of a webcrawler that would allow for the constant update of this dataset periodically enriching it with the current information. This updated MAEC dataset offers a wider range of company conversations from recent years while maintaining the integrity and quality of the original collection. Essentially, the MAEC dataset received a modern update, solidifying its position as a premier tool for analyzing company sentiments and behaviors.

The journey through the realm of sentiment analysis in this study brought forth several insights. The blend of CNN and LSTM architectures to decode sentiments demonstrated a significant evolution in understanding the nuances of sentiments in textual data. The baseline was established by training it on the Twitter dataset. This allowed for the creation of a sentiment analysis tool that was capable of labeling the expansive textual part of the MAEC dataset. Thereby, providing insights for the involvement of sentiment analysis with finance.

Transitioning from the Twitter dataset to the MAEC dataset was not just an exercise in model adaptability but also a deep dive into the world of financial linguistics.

The difference in language style and content structure between tweets and earnings calls underscored the essence of dataset-specific preprocessing. It also highlighted the model's capability to capture sentiments in varying contexts, be it the informal diction of tweets or the structured jargon of earnings calls. This ultimately led to the labeling of the sentiments of the textual portion of the MAEC dataset, thus, providing it a new edge.

The model's performance on the MAEC dataset has profound implications. Financial experts and stock market enthusiasts can harness these sentiment labels to better understand the underlying mood of an earnings call, potentially providing an edge in stock market predictions and investment decisions.

5.1 Future Work

The vast landscape of sentiment analysis presents numerous avenues for future exploration. One of the most promising avenues lies within the MAEC dataset itself. Its unique text-audio paired structure provides an opportunity for exploring multi-modal sentiment analysis. Incorporating vocal tonality and pace into sentiment calculations might offer a more comprehensive sentiment assessment.

An intriguing proposition would be to delve into temporal sentiment analysis. With an expansive dataset like MAEC, tracking sentiment trends over the years could reveal patterns. Such patterns might resonate with economic shifts, pivotal industry innovations, or even broader global events, offering a sentiment lens to view financial history.

Future researchers might want to explore different architectures, maybe transformers like BERT, to understand their efficacy in this domain. The precedent set by training on the Twitter dataset makes a compelling case for further ventures into transfer learning. There's potential in gauging the effectiveness of models pre-trained on various textual datasets when applied to financial transcripts.

As the influence of sentiment analysis grows, especially in domains as impactful as finance, ethical considerations must be paramount. Ensuring that models are free from biases and that data handling respects privacy norms will be crucial. In the practical realm, envisioning a platform where analysts could get real-time sentiment assessments of financial transcripts can be an exciting direction, bringing the power of sentiment analytics to the fingertips of financial analysts. This research illuminated the path of sentiment analysis in financial contexts, by adding sentiments to the earnings calls and exploring the possibility of the intersection of the finance and human sentiments.

Bibliography

[1] A. Kuznetsova, H. Rom, N. Alldrin, J. R. R. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari, "The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale," *CoRR*, vol. abs/1811.00982, 2018. [Online]. Available: http://arxiv.org/abs/1811.00982

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: visual question answering," *CoRR*, vol. abs/1505.00468, 2015. [Online]. Available: http://arxiv.org/abs/1505.00468

[3] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," 2016. [Online]. Available: https://arxiv.org/abs/1612.06890

[4] Q.-T. Truong and H. Lauw, "Multimodal review generation for recommender systems." New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3308558.3313463

[5] a. a. KazAnova, "Sentiment140 dataset with 1.6 million tweets," Sep 2017. [Online]. Available: https://www.kaggle.com/datasets/kazanova/sentiment140

[6] D. Liu, Y. Cui, W. Tan, and Y. Chen, "Sg-net: Spatial granularity network for one-stage video instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9816–9825.

[7] Y. Cui, L. Yan, Z. Cao, and D. Liu, "Tf-blender: Temporal feature blender for video object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8138–8147.

[8] R. Khan, "Importance of datasets in machine learning and ai research," *DataToBiz*, Nov 2022. [Online]. Available: https://www.datatobiz.com/blog/datasets-in-machine-learning/

[9] L. Cao, "Ai in finance: Challenges, techniques and opportunities," 2021.

[10] A. Prasad and A. Seetharaman, "Importance of machine learning in making investment decision in stock market," *Vikalpa*, vol. 46, no. 4, pp. 209–222, 2021. [Online]. Available: https://doi.org/10.1177/02560909211059992

[11] D. Liu, Y. Cui, L. Yan, C. Mousas, B. Yang, and Y. Chen, "Densernet: Weakly supervised visual localization using multi-scale feature aggregation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 7, 2021, pp. 6101–6109.

[12] G. Fatouros, J. Soldatos, K. Kouroumali, G. Makridis, and D. Kyriazis, "Transforming sentiment analysis in the financial domain with chatgpt," *Machine Learning with Applications*, vol. 14, p. 100508, Dec. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.mlwa.2023.100508

[13] C. S. Throckmorton, W. J. Mayew, M. Venkatachalam, and L. M. Collins, "Financial fraud detection using vocal, linguistic and financial cues," *Decision Support Systems*, vol. 74, p. 78–87, 2015.

[14] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," 2019.

[15] J. Li, L. Yang, B. Smyth, and R. Dong, "Maec," *Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management*, 2020.

[16] B. Zhang, H. Yang, T. Zhou, A. Babar, and X.-Y. Liu, "Enhancing financial sentiment analysis via retrieval augmented large language models," 2023.

[17] W. Lakhchini, R. Wahabi, and M. Kabbouri, "Artificial intelligence machine learning in finance: A literature review," 12 2022.

[18] S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix, "Identification of fraudulent financial statements using linguistic credibility analysis," *Decision Support Systems*, vol. 50, no. 3, pp. 585–594, 2011, on quantitative methods for detection of financial fraud. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923610001338

[19] L. N. Driscoll, "A validity assessment of written statements from suspects in criminal investigations using the scan technique," *Police Stud.: Int'l Rev. Police Dev.*, vol. 17, p. 77, 1994.

[20] P. M. Dechow, R. G. Sloan, and A. P. Sweeney, "Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the sec," *Contemporary accounting research*, vol. 13, no. 1, pp. 1–36, 1996.

[21] D. F. LARCKER and A. A. ZAKOLYUKINA, "Detecting deceptive discussions in conference calls," *Journal of Accounting Research*, vol. 50, no. 2, p. 495–540, 2012.

[22] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction," in *International Joint Conference on Artificial Intelligence*, 2015.

[23] "Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction," *CoRR*, vol. abs/1712.02136, 2017, withdrawn. [Online]. Available: http://arxiv.org/abs/1712.02136

[24] X. Zhang, Y. Zhang, S. Wang, Y. Yao, B. Fang, and P. S. Yu, "Improving stock market prediction via heterogeneous information fusion," *Knowledge-Based Systems*, vol. 143, pp. 236–247, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705117306032

[25] F. G. D. C. Ferreira, A. H. Gandomi, and R. T. N. Cardoso, "Artificial intelligence applied to stock market trading: A review," *IEEE Access*, vol. 9, pp. 30 898–30 917, 2021.

[26] S. Mokhtari, K. K. Yen, and J. Liu, "Effectiveness of artificial intelligence in stock market prediction based on machine learning," *International Journal of Computer Applications*, vol. 183, no. 7, p. 1–8, Jun. 2021. [Online]. Available: http://dx.doi.org/10.5120/ijca2021921347

[27] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," 2019. [Online]. Available: https://arxiv.org/abs/1908.05787

[28] Y. Qin and Y. Yang, "What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 390–401. [Online]. Available: https://aclanthology.org/P19-1038

[29] W. Wang, C. Han, T. Zhou, and D. Liu, "Visual recognition with deep nearest centroids," *ICLR*, 2022.

[30] G. Meena, K. K. Mohbey, and S. Kumar, "Sentiment analysis on images using convolutional neural networks based inception-v3 transfer learning approach," *International Journal of Information Management Data Insights*, vol. 3, no. 1, p. 100174, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667096823000216

[31] A. Wang, A. Liu, R. Zhang, A. Kleiman, L. Kim, D. Zhao, I. Shirai, A. Narayanan, and O. Russakovsky, "Revise: A tool for measuring and mitigating bias in visual datasets," 2020. [Online]. Available: https://arxiv.org/abs/2004.07999

[32] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[33] M. F. Ishmam, M. S. H. Shovon, M. F. Mridha, and N. Dey, "From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities," 2023.

[34] J. W. Cho, D.-j. Kim, H. Ryu, and I. S. Kweon, "Generative bias for visual question answering," 2022. [Online]. Available: https://arxiv.org/abs/2208.00690

[35] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," *CoRR*, vol. abs/1712.00377, 2017. [Online]. Available: http://arxiv.org/abs/1712.00377

[36] C. Kervadec, G. Antipov, M. Baccouche, and C. Wolf, "Roses are red, violets are blue... but should vqa expect them to?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 2776–2785.

[37] R. Cadene, C. Dancette, H. Ben younes, M. Cord, and D. Parikh, "Rubi: Reducing unimodal biases for visual question answering," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/51d92be1c60d1db1d2e5e7a07da55b26-Paper.pdf

[38] Y. Lu, Q. Wang, S. Ma, T. Geng, Y. V. Chen, H. Chen, and D. Liu, "Transflow: Transformer as flow learner," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 063–18 073.

[39] Q. Wang, L. Yang, X. Quan, F. Feng, D. Liu, Z. Xu, S. Wang, and H. Ma, "Learning to generate question by asking question: a primal-dual approach with uncommon word generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 46–61.

[40] Z. Cao, Z. Chu, D. Liu, and Y. Chen, "A vector-based representation to enhance head pose estimation," in *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, 2021, pp. 1188–1197.

[41] Z. Cheng, J. Liang, H. Choi, G. Tao, Z. Cao, D. Liu, and X. Zhang, "Physical attack on monocular depth estimation with optimal adversarial patches," in *European Conference on Computer Vision*. Springer, 2022, pp. 514–532.

[42] Q. Wang, Y. Fang, A. Ravula, F. Feng, X. Quan, and D. Liu, "Webformer: The web-page transformer for structure information extraction," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3124–3133.

[43] D. Liu, J. Liang, T. Geng, A. Loui, and T. Zhou, "Tripartite feature enhanced pyramid network for dense prediction," *IEEE Transactions on Image Processing*, 2023.

[44] Z. Cheng, J. Liang, G. Tao, D. Liu, and X. Zhang, "Adversarial training of self-supervised monocular depth estimation against physical-world attacks," *ICLR*, 2023.

[45] C. Han, Q. Wang, Y. Cui, Z. Cao, W. Wang, S. Qi, and D. Liu, "Eˆ 2vpt: An effective and efficient approach for visual prompt tuning," *ICCV*, 2023.

[46] Z. Cao, D. Liu, Q. Wang, and Y. Chen, "Towards unbiased label distribution learning for facial pose estimation using anisotropic spherical gaussian," in *European Conference on Computer Vision*. Springer, 2022, pp. 737–753.

[47] Z. Qin, X. Lu, D. Liu, X. Nie, Y. Yin, J. Shen, and A. C. Loui, "Reformulating graph kernels for self-supervised space-time correspondence learning," *IEEE Transactions on Image Processing*, 2023.

[48] L. Yan, C. Han, Z. Xu, D. Liu, and Q. Wang, "Prompt learns prompt: exploring knowledge-aware generative prompt collaboration for video captioning," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI). International Joint Conferences on Artificial Intelligence Organization. https://doi. org/10.24963/ijcai*, vol. 180, 2023.

[49] J. W. Goodell, S. Kumar, W. M. Lim, and D. Pattnaik, "Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis," *Journal of Behavioral and Experimental Finance*, vol. 32, p. 100577, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214635021001210

[50] J. L. C. Sanz and Y. Zhu, "Toward scalable artificial intelligence in finance," in *2021 IEEE International Conference on Services Computing (SCC)*, 2021, pp. 460–469.

[51] J. Maqbool, P. Aggarwal, R. Kaur, A. Mittal, and I. A. Ganaie, "Stock prediction by integrating sentiment scores of financial news and mlp-regressor: A machine learning approach," *Procedia Computer Science*, vol. 218, pp. 1067–1078, 2023, international Conference on Machine Learning and Data Engineering. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923000868

[52] F. Ferreira, A. Gandomi, and R. Cardoso, "Artificial intelligence applied to stock market trading: A review," *IEEE Access*, vol. PP, pp. 1–1, 02 2021.

[53] P. Cheeseman and R. Oldford, *Selecting Models from Data: Artificial Intelligence and Statistics IV*, 01 1994, vol. 89.

[54] Y. J. Cruz, M. Rivas, R. Quiza, R. E. Haber, F. Castaño, and A. Villalonga, "A two-step machine learning approach for dynamic model selection: A case study on a micro milling process," *Computers in Industry*, vol. 143, p. 103764, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166361522001610

[55] I. Gupta, T. K. Madan, S. Singh, and A. K. Singh, "Hisa-smfm: Historical and sentiment analysis based stock market forecasting model," 2022.

[56] J. X. Zhao, Y. Xie, K. Kawaguchi, J. He, and M. Q. Xie, "Automatic model selection with large language models for reasoning," 2023.

[57] L. Cao, "Ai in finance: Challenges, techniques, and opportunities," vol. 55, no. 3, 2022. [Online]. Available: https://doi.org/10.1145/3502289

[58] J. Liang, T. Zhou, D. Liu, and W. Wang, "Clustseg: Clustering for universal segmentation," *ICML*, 2023.

[59] S. Liu, C. Zhang, and J. Ma, "Cnn-lstm neural network model for quantitative strategy analysis in stock markets," 10 2017, pp. 198–206.

[60] J. Li, L. Yang, B. Smyth, and R. Dong, "Maec: A multimodal aligned earnings conference call dataset for financial risk prediction," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, ser. CIKM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 3063–3070. [Online]. Available: https://doi.org/10.1145/3340531.3412879

[61] H. Yang, Y. Zhao, J. Liu, Y. Wu, and B. Qin, "Macsa: A multimodal aspect-category sentiment analysis dataset with multimodal fine-grained aligned annotations," 2022.

[62] A. A, R. R, V. R. S, and A. M. Bagde, "Predicting stock market time-series data using cnn-lstm neural network model," 2023.

[63] Z. Shi, Y. Hu, G. Mo, and J. Wu, "Attention-based cnn-lstm and xgboost hybrid model for stock prediction," 2023.

[64] C. Huang, Z. Zhang, B. Mao, and X. Yao, "An overview of artificial intelligence ethics," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 4, pp. 799–819, 2023.

[65] A. A. Khan, S. Badshah, P. Liang, B. Khan, M. Waseem, M. Niazi, and M. A. Akbar, "Ethics of ai: A systematic literature review of principles and challenges," 2021.

[66] L. P. Silvestrin, H. van Zanten, M. Hoogendoorn, and G. Koole, "Transfer learning across datasets with different input dimensions: An algorithm and analysis for the linear regression case," *Journal of Computational Mathematics and Data Science*, vol. 9, p. 100086, Dec. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.jcmds.2023.100086

[67] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017.

[68] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen, "Image data augmentation for deep learning: A survey," 2023.

[69] K. L. Tan, C. P. Lee, and K. M. Lim, "A survey of sentiment analysis: Approaches, datasets, and future research," *Applied Sciences*, vol. 13, no. 7, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/7/4550

[70] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," *CoRR*, vol. abs/1811.12808, 2018. [Online]. Available: http://arxiv.org/abs/1811.12808

[71] M. Birjali, M. Kasri, and A. beni hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, p. 107134, 05 2021.

[72] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9603–9611, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417415005126