Multimodal Multi-Speaker Merger & Acquisition Financial Modeling: A New Task, Dataset, and Neural Baselines

Ramit Sawhney[†*], **Mihir Goyal**[†*], **Prakhar Goel**[†*], **Puneet Mathur**[‡], **Rajiv Ratn Shah**[†]
† Department of Computer Engineering, IIIT Delhi
‡ University of Maryland, College Park
{ramits, mihir17166, prakhar17306, rajivratn}@iiitd.ac.in
[‡]puneetm@umd.edu

Abstract

Risk prediction is an essential task in financial markets. Merger and Acquisition (M&A) calls provide key insights into the claims made by company executives about the restructuring of the financial firms. Extracting vocal and textual cues from M&A calls can help model the risk associated with such financial activities. To aid the analysis of M&A calls, we curate a dataset of conference call transcripts and their corresponding audio recordings for the time period ranging from 2016 to 2020. We introduce `M3ANet`, a baseline architecture that takes advantage of the multimodal multi-speaker input to forecast the financial risk associated with the M&A calls. Empirical results prove that the task is challenging, with the proposed architecture performing marginally better than strong BERT-based baselines. We release the `M3A` dataset and benchmark models to motivate future research on this challenging problem domain.

1 Introduction

Mergers and Acquisitions (M&As)[1] conference calls are events preceding financial transactions involving two or more entities such that either one of the participant companies takes over the other(s) and establishes itself as the owner (termed as "acquisition") or when one company combines with another to become a joint entity (termed as "merger"). In these M&A conference calls, the participating companies' management makes a presentation to the call participants, such as market analysts, media personnel, and other stakeholders, explaining the rationale for the deal and possible roadblocks to deal completion (Dasgupta et al., 2020). Following the presentation segment, there is a Q&A segment in which the call participants ask questions to which the management responds.

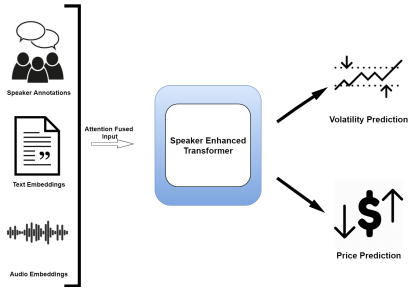


Figure 1: A schematic of our proposed approach (`M3A`) that leverages three types of input modalities: text utterances from the call transcripts, audio clips, and speaker specific input, for financial modeling tasks.

Building on the important information that M&As provide, academic research, the financial press, and other media give a great deal of attention. One of these discussions' principal aspects lies in how the deals may affect the company's valuation (Moeller et al., 2003; Fraunhoffer et al., 2018) and future growth. A significant focus in financial and economic literature has been on understanding whether M&As create or destroy value. Consequently, shareholders critically analyze the deals to estimate the potential stock price and stock price volatility post the M&A conference call.

Identifying the gap in natural language processing (NLP) literature on the lack of resources to study M&A conference calls with their text transcripts and audio recordings, we take the first step in multimodal financial modeling in the M&A space. Such data can allow academicians to study M&A calls further, especially with the rich multimodal data. It shall enable studies that focus not only on the words spoken in the call but also in the manner they were spoken, a relatively unexplored field in financial forecasting, as shown in Figure 1.

A salient aspect of conference calls is that, unlike text reports, the company's management interacts with external stakeholders and asks questions. This

[*] Equal contribution
[1] https://www.investopedia.com/mergers-and-acquisitions

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 6751–6762
August 1–6, 2021.

Financial Analyst: So, from the beginning, you've always made a point that you thought you had a better regulatory path because of **Comcast** owning distribution. But **I guess** following **last week's Judge Leon's decision, do you think** the path for both Comcast and Disney are the same, that's what we heard **last week** from Comcast. So how would you defend **I guess** the argument that both regulatory paths look similar? **And then** secondly, any update on **Sky**. **Christine's analysis** puts Sky in terms of **a if, or maybe or maybe not** we get it, but what's the latest thinking on Sky and the following bid on Sky?

Disney CEO: Well, I'll answer the second part first, Michael. We don't have an update on **Sky**. **Fox** is in the driver's seat regarding the 61% that they do not control. You've read the **recent** news about the regulatory path that they have and I won't say more than that. **Let** me say a few things about the regulatory process. I mentioned in my remarks that we're confident and that we have a clear path. **As you know**, issues aside, we've been working with regulatory authorities not just in the **United States but** in jurisdictions across the world now for **six months** and we've made a lot of progress towards obtaining the regulatory approvals that are necessary, **I think** that includes supplying regulatory authorities with a tremendous amount of information, **basically** heeding their requests for all sorts of documentation regarding this acquisition and the potential impact that it has on markets across the world.

Figure 2: M&A calls have a Q&A session where financial stakeholders can ask questions to the company executives. In such sessions, company executives have to be impromptu with their responses, allowing informal words to seep in. This example Q&A session is from the call regarding the acquisition of 21st Century Fox by Disney, dated June 20, 2018. In the example, an analyst poses a few questions to the company executives (depicted in **yellow**). The CEO of Disney responds to these questions, where we notice some cases of informal speech (depicted in **purple**). The Executive's response however mainly focused on specific objects or entities (depicted in **red**) intermixed with some time-based information (depicted in **green**).

interaction presents an opportunity of analyzing not just the management's claims but also the way they express them. In Figure 2, we highlight the various components in a short Q&A interaction. Often, both the transcript and the audios of the calls are available to the public.

Vocal cues play a critical role in verbal communication as they can provide support or discredit the verbal message that is being spoken (Jiang and Pell, 2017). For example, consider if the CEO of the acquiring company exhibits confidence in the statement - "we are confident that this acquisition will bring us profits," however, displays nervousness while justifying technical details of the deal, we may infer contradiction in the claims of a successful M&A. Vocal cues have been proven indicators of emotions like deceit and nervousness (Belin et al., 2017; Sporer and Schwandt, 2006). Past research (Qin and Yang, 2019; Sawhney et al., 2020c) shows that the addition of vocal cues has helped with the task of financial predictions and enrich the learned representations.

Our contributions can be summarized as:

- We curate a public dataset `M3A`[2] (Multimodal Multi-Speaker Merger & Acquisition Call Financial Forecasting Dataset) that consists of 816 M&A conference calls spanning over 545 hours between 2016 to 2020 with their transcripts and audio recordings, segmented by utterances and aligned with the audio.

- We accompany the dataset with neural baseline architectures that use the multimodal multi-speaker input to predict stock volatility and price movement.

- To the best of our knowledge, no such M&A conference call dataset exists in academia, and our proposed methodology, `M3ANet` is the first deep learning approach for financial predictions on M&A conference calls.

2 Related Work

M&A Conference Calls Financial reports and conference calls have been shown to have a correlation with the stock market and improve financial predictions (Bowen et al., 2001; Kogan et al., 2009). Studies have also been carried out specifically for M&A calls, showing their effect on the market (Dasgupta et al., 2020; Hu et al., 2018). However, there exists a gap in leveraging neural predictive modeling on using verbal and vocal cues pertaining to M&A calls for financial forecasting.

[2]The source code, processed features, and details on acquiring raw data are available at https://github.com/midas-research/m3a-acl

Financial Forecasting Research has shown historical pricing data to be useful in predicting financial risk modeling (Kristjanpoller et al., 2014; Zheng et al., 2019; Dumas et al., 2009). It also considers volatility as an indicator of uncertainty, which helps make decisions regarding investments (Heston, 1993; Johnson and Shanno, 1987; Scott, 1987). Previous work often use numerical features (Liu and Chen, 2019; Nikou et al., 2019) in approaches like neural networks (Kim et al., 2019; Luo et al., 2017), graph neural networks (Sawhney et al., 2020b), and time-series models (Bollerslev, 1986; Engle, 1981). On the other hand, we are interested in analyzing multimodal data like text and audio, which can hold completely different information for predictive models.

Natural Language Processing and Finance For any system using human interactions to determine financial risk or stock movements, it is necessary to determine the relationship between the various words to determine the speaker's sentiment. Advances in NLP have been utilized in many approaches to show financial information significantly improving performance in forecasting tasks like volatility and stock price prediction (Wang et al., 2013; Ding et al., 2015; Mittermayer and Knolmayer, 2007). Research has also shown that social media affects the stock market (Bollen et al., 2010; Oliveira et al., 2017; Sawhney et al., 2020a). Machine learning methods using simple bag-of-words features to represent the financial documents used in previous research (Kogan et al., 2009; Rekabsaz et al., 2017) largely ignore the inter-dependencies between the sentences. To fill the gap, recent approaches have moved towards newer models such as transformers (Yang et al., 2020) and reinforcement learning (Sawhney et al., 2021b) over natural language data for financial forecasting.

Multimodality and Financial Forecasting Research shows that psychological and behavioral elements are often indicators of stock price movement (Malkiel, 2003). Vocal cues have been proven effective in portraying these elements (Wurm et al., 2010; Hobson et al., 2011; Jiang and Pell, 2017). Thus, it is no surprise that multimodal architectures that use these cues for financial predictions have seen significant improvements in their performances (Yang et al., 2020; Sawhney et al., 2020d).

Speaker Context Encoding Past research (Zhang et al., 2019; Li et al., 2020) in fields like emotion recognition have seen the improved performance on their prediction tasks with the addition of speaker context. Models with data related to spoken text benefit when the input is enriched with information about who spoke what.

3 Problem Formulation

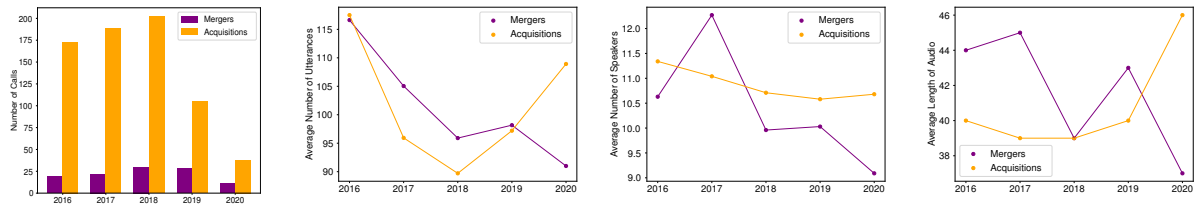
Consider an M&A call $\chi \in \{\chi_1, \chi_2, \ldots, \chi_M\}$, which comprises multimodal components: $\chi = [t; a]$. Here, t is the sequence of textual utterances (sentences)[3] of the call transcript and can be represented as $[t_1, t_2 ... t_N]$ where t_i is the i^{th} utterance of the call and N is the maximum number of utterances in any call. Similarly, a is the sequence of corresponding call audios for the textual utterances (sentences) and can be represented as $[a_1, a_2 ... a_N]$ where a_i is the i^{th} call audio. The call's utterances are annotated with speaker information $s = [s_1, s_2 ... s_N]$, where s_i is the speaker of the i^{th} utterance and where each speaker in the call may have spoken one or more utterances. Each M&A conference call may have two or more participating companies, with at least one publicly-traded company with publicly available stock price information. We limit the scope of the problem being solved by forecasting predictions for just one of the participant companies with the larger market valuation (in case of a merger) or the acquiring company (in case of an acquisition). We now describe the two prediction tasks that we utilize to train `M3ANet` on.

Measuring stock volatility Following (Kogan et al., 2009), we formulate stock volatility as a regression problem. For a given stock with a close price of p_k on the trading day k, we calculate the average log volatility as the natural log of the standard deviation of return prices r in a window of τ days as:

$$v_{[0,\tau]} = \ln\left(\sqrt{\frac{\sum_{k=1}^{\tau}(r_k - \bar{r})^2}{\tau}}\right) \quad (1)$$

where $r_k = \frac{p_k - p_{k-1}}{p_{k-1}}$ is the return price on day k for a given stock, and $\bar{r}$ is the average return price over a period of τ days.

[3] We restrict the scope of segmentation to a sentence level as opposed to a more granular level such as the word level owing to the higher complexity and noise involved in word-level segmentation for long M&A calls.



(a) Yearly frequency of calls. (b) Mean # of utterances. (c) Mean # of speakers. (d) Mean audio length.

Figure 3: Statistics pertaining to the M3A dataset across modalities, types of calls, and years.

Formalizing price movement prediction Following (Xu and Cohen, 2018), we define price movement $y_{d-\tau,d}$ over a period of τ days as a binary classification task. For a given stock, we employ its close price, which can either rise or fall on a day d compared to a previous day $d-\tau$, to formulate the classification task as:

$$y_{[d-\tau,d]} = \begin{cases} 1, p_{d+\tau} > p_d, \\ 0, p_{d+\tau} \leq p_d \end{cases} \quad (2)$$

Given an acquisition conference call χ, our learning objective is to predict the average negative log volatility $v_{[0,\tau]}$ and price movement $y_{[d-\tau,d]}$ using the conference call data $\chi = [t; a]$.

4 Curating M3A: Dataset Creation

4.1 Data Acquisition

We curate our dataset, M3A, by acquiring audio records and text transcripts from the Bloomberg Terminal.[4] Since the conference calls were reliably available from 2016, we filter and list all M&A calls between 2016 and 2020. To limit the scope, we ensured the calls were in English, had their domicile as the U.S.A., and had 'merger' or 'acquisition' in their title. The Bloomberg Terminal often only provides the stock ticker for the acquiring company (in case of an acquisition) and the company with a more prominent marker valuation (in case of a merger). To maintain uniformity, we decide only to use the given stock information. We pull the adjusted closing price data from Yahoo Finance.[5]

The dataset comprises 816 conference calls. The mean number of speakers across the calls is 10.68 ± 4.17, with a maximum of 31 speakers. The mean number of utterances across the calls is 100.54 ± 38.32 utterances and a maximum of 284 utterances in a call. The mean length comes out to be 40.15 ± 15.15 minutes and a maximum length of 98.15 minutes for the audio clips. We provide further statistics in Figure 3. Looking at year-wise trends, we see that acquisitions are consistently more frequent that mergers every year. Further, we note that mergers see a decreasing trend in the number of utterances and acquisitions have a consistent number of speakers in M&A calls. We also note that acquisitions conference calls seem to be increasing in length as the years progress.

We chronologically divide our dataset into a train, validation, and test set in the ratio of 70 : 10 : 20, respectively. Such a split ensures that future data is not used for forecasting past data.

4.2 Call Segmentation and Alignment

Each transcript of the dataset begins with the company's details with the larger market valuation (in case of a merger) or the acquiring company (in case of an acquisition). These details include the company's name, stock ticker, and the date of the call. The transcript then lists the speakers in the call and their position in the companies, if any. The call contents follow the list of speakers. The contents are separated by utterances and are annotated with the utterances' speakers.

Given our dataset, we have the option to choose between transcript-level, utterance-level, and word-level embeddings. We decide to use utterance-level embeddings.[6] We select utterances with at least ten words to ensure better parsing of the transcript and parse the texts to extract all valid utterances.

Since we are working with audio files, it becomes essential that we can segment them such that we can align them with their corresponding utterances in the text transcript. To achieve this alignment, we have used the Aeneas[7] library to per-

[4] https://bba.bloomberg.net/

[5] https://in.finance.yahoo.com/

[6] Transcript-level embeddings are too coarse for our task. We experimented with word-level embeddings but found that the performance degraded.

[7] https://www.readbeyond.it/aeneas/

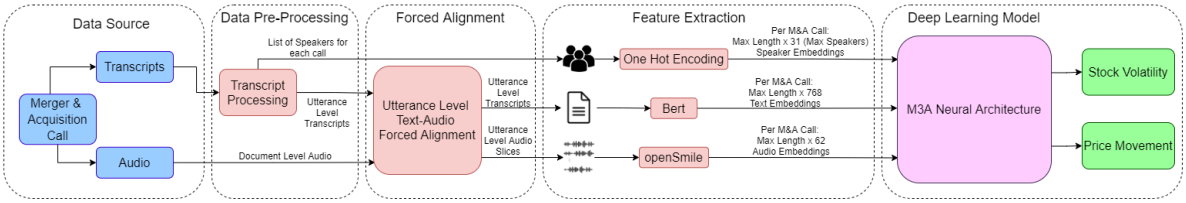


Figure 4: Data Pipeline: An overview of the processing involved with each data point including segmentation, encoding of modalities, speaker information augmentation and prediction.

form the forced alignment. The Forced Alignment algorithm takes as input a text file divided into fragments and an unfragmented audio file. It processes the input to output a synchronization map, which automatically associates a time interval in the audio file to its corresponding text fragment. Aeneas uses the Sakoe-Chiba Band Dynamic Time Warping (DTW) (Sakoe and Chiba, 1978) forced alignment algorithm, which has been proven to improve discrimination between words and has superior performance over other conventional algorithms.

5 Methodology

5.1 Text and Audio Encoding

Text Encoding We compute an utterance's textual encoding as the arithmetic mean of all its word vectors. BERT is well known as an effective pre-trained language-based model for extracting word-embeddings (Biswas et al., 2020) for a variety of language modeling tasks. We use Uncased Base BERT (Devlin et al., 2019) to extract the word embeddings. For each call, we represent the text utterances as $[t_1, t_2, \ldots, t_N]$. As seen from Figure 4, we embed each text utterance t_i to get its corresponding 768-dimensional text encoding g_i using BERT such that $g_i = \texttt{BERT}(t_i)$ for each $i \in [1, N]$.

Audio Encoding We use the OpenSMILE[8] library to extract the audio features at a sampling rate of 10ms and choose the set of 62 geMAPS features described in (Eyben et al., 2016). This set includes features like pitch, jitter, loudness, etc., which have proven to be effective in audio analysis tasks (Chao et al., 2015). For each call, we represent the audio clips of the utterances as $[a_1, a_2, \ldots, a_N]$. We embed each audio utterance a_i to its corresponding 62-dimensional encoding h_i using OpenSMILE such that $h_i = \texttt{OpenSMILE}(a_i)$ for each $i \in [1, N]$.

[8] https://pypi.org/project/opensmile/

Motivation for Speaker Information Infusion The audio encodings help decipher the vocal cues in the text transcript's context to support or discredit the speaker's claims. However, it is critical for the system to recognize the importance of the utterance's speaker to gauge its impact on financial predictions. This requires the information about the speaker of each utterance to be augmented to the input. Prior research (Zhang et al., 2019; Li et al., 2020) shows the addition of speaker context helps improve prediction performance on tasks involving datasets with spoken texts.

M&A calls have utterances spoken by the company's management (the decision-making force of the company), by analysts (who want to gauge the risk in the company's decisions), or even just the operator (often an impartial person). Capturing this speaker context will allow us to decide how much impact a specific utterance can have on a company's stock price. Thus, we extract the speaker information for each utterance. We parse the list of speakers from the transcripts and assign an ID to each of the speakers. The IDs start from 1 and are assigned incrementally to each speaker in the order in which they are listed. The operator of the call is assigned the ID 0. We then annotate each of the utterances based on who spoke it. Finally, we use one-hot encoding to represent the speaker encoding s of each utterance in the call.

5.2 `M3ANet`: Speaker Transformer

The Transformer (Vaswani et al., 2017) uses multi-head attention and position embeddings to learn the relationship between different utterances. The multimodal input requires the model to learn the inter-dependencies between the audio and the text features. `M3ANet` can then use the audio cues to affirm or discredit the spoken message and make an informed prediction. The idea behind `M3ANet` is to use attention to weigh the importance of each modality at different timestamps. We then aug-

ment the data with the speaker encoding and allow the Transformer to extract the multimodal interdependencies for performing the prediction tasks.

Attention-Fusion Before we can fuse the inputs, we need to linearly transform the text embeddings to ensure the multimodal embeddings' sizes are the same. We then extract the attention weights to calculate the attended inputs similar to (Hori et al., 2017). These attention weights describe the importance of a specific modality concerning the other modality. We multiply the text and audio features by their attention weights W_T and W_A respectively to get the attended input, followed by fusing them. The following equations formalize the attention mechanism used:

$$W_T = \texttt{softmax}(gW_{wt} + b_{wt}) \quad (3)$$

$$W_A = \texttt{softmax}(hW_{wa} + b_{wa}) \quad (4)$$

$$W_T = \frac{W_T}{W_T + W_A}, W_A = \frac{W_A}{W_T + W_A} \quad (5)$$

$$X_{fused} = gW_T + hW_A \quad (6)$$

where W_{wt} and b_{wt} represent the text attention layer, W_{wa} and b_{wa} represent the audio attention layer and $+$ represents addition.

Sentence-Level Transformer To model the sequence of textual and audio embeddings of the M&A calls, we augment the fused multimodal embeddings X_{fused} with position embeddings pos by addition and the speaker information by concatenation (represented by $\oplus$). pos has the same dimensions as X_{fused}, $pos_{j,ind}$ represents the value of the positional embedding for the j^{th} utterance at index ind. The augmentation is summarised as follows:

$$pos_{j,2l}, pos_{j,2l+1} = \sin\left(\frac{j}{10^{\frac{8l}{d}}}\right), \cos\left(\frac{j}{10^{\frac{8l}{d}}}\right) \quad (7)$$

$$X_{final} = (X_{fused} + pos) \oplus s \quad (8)$$

The Transformer block uses the augmented feature set for further processing, following which the intermediate tensors are passed through two consecutive dense layers to output the task prediction as follows:

$$O_1 = ReLU(W_{l1}I_1 + b_{l1}) \quad (9)$$

$$y = \sigma(W_{l2}O_1 + b_{l2}) \quad (10)$$

where, W_{l1} and b_{l1} represent the first linear layer, W_{l2} and b_{l2} represent the second linear layer, I_1 and O_1 represent the input to the first and second dense layer after being passed through the sentence transformer, while σ represents the final activation function and y represents the final prediction from the activation corresponding to the task. We use ReLU for the final prediction in the volatility prediction task and sigmoid for the price prediction task. We then use Mean Squared Error (MSE) and Binary Cross-Entropy (BCE) losses to train the output for volatility prediction and stock price movement prediction, respectively.

6 Experimental Setup

6.1 Baselines

We compare `M3ANet` against modern baselines across modalities for both the tasks. We employ GloVe (Pennington et al., 2014), FinBERT (Araci, 2019) and Roberta (Liu et al., 2019) to embed the text and choose an LSTM + Dense layer architecture as a benchmark for both volatility and price movement prediction. We also use all three (text, audio, and multimodal) variants of the Multimodal Deep Regression Model (MDRM) (Qin and Yang, 2019) as baselines.

6.2 Training Setup

We tune `M3ANet`'s hyper-parameters using Grid Search. We summarize the range of hyperparameters tuned on: size of the transformer's feed-forward layer and size of the linear layers $\in \{16, 32, 64\}$, dropout $\delta \in \{0.0, 0.1, 0.25, 0.5\}$, batch size $b \in \{32, 64, 128\}$ and learning rate $e \in \{0.1, 0.01, 0.001, 0.0001\}$. The experiment results in the following optimal choices of the hyper-parameters: $b = 64$, $e = 0.001$, feed forward network size (Volatility) $= 16$, hidden layer size (Volatility) $= 16$ and δ (Volatility) $= 0.1$, , feed forward network size (Movement) $= 64$, hidden layer size (Movement) $= 32$, δ (Movement) $= 0.0$.

We implement all methods with Keras[9] and Google Colab.[10], using ReLU as our hidden layer activation function and optimize using Adam. We choose the highest performing model during the training phase on our validation set and chosen evaluation metrics as our best model. We zero-pad the calls that have less than the maximum number of utterances/speakers for efficient batching. We experiment with trading periods $\tau \in \{3, 7, 15\}$

[9] https://keras.io/

[10] https://research.google.com/colaboratory/

Model	Volatility Prediction			Price Prediction					
	MSE_3	MSE_7	MSE_{15}	$F1_3$	$F1_7$	$F1_{15}$	MCC_3	MCC_7	MCC_{15}
RoBERTa + LSTM	0.78 (0.009)	0.58 (0.009)	0.47 (0.006)	0.57	0.58	0.49	0.19	0.22	0.10
GloVe + LSTM	0.80 (0.005)	0.60 (0.004)	0.48 (0.005)	0.55	0.56	0.42	0.19	0.22	0.02
FinBERT + LSTM	0.78 (0.008)	0.60 (0.004)	0.47 (0.005)	0.58	0.58	0.48	0.20	0.21	0.06
MDRM (T)	0.79 (0.003)	0.59 (0.003)	0.47 (0.002)	0.58	0.56	0.48	**0.20**	**0.19**	0.12
MDRM (A)	0.79 (0.004)	0.60 (0.002)	0.47 (0.003)	0.24	0.36	0.12	0.02	0.17	0.00
MDRM (T+A)	0.78 (0.005)	0.58 (0.003)	0.46 (0.002)	**0.59**	0.58	0.46	0.19	**0.19**	0.11
M3ANet (Ours)	**0.77 (0.018)***	**0.57 (0.016)***	**0.46 (0.011)***	**0.59**	**0.59**	**0.50***	0.19	**0.19**	**0.13**

Table 1: Mean τ-day volatility MSE and price movement prediction results (mean and stdev. of 5 runs for each approach). * indicates that the result is significantly better than the MDRM (T+A). **Bold** denotes best performance.

Model	Volatility Prediction			Price Prediction					
	MSE_3	MSE_7	MSE_{15}	$F1_3$	$F1_7$	$F1_{15}$	MCC_3	MCC_7	MCC_{15}
Transformer (T)	0.79 (0.0130)	0.62 (0.0310)	0.47 (0.004)	0.50	0.54	0.40	0.13	0.16	0.11
Transformer (A)	0.82 (0.0180)	0.61 (0.0140)	0.49 (0.013)	0.53	0.59	0.50	0.13	0.18	0.13
Transformer (T+A: Concat)	0.80 (0.0006)	0.61 (0.0006)	0.48 (0.0003)	0.09	0.16	0.06	0.00	0.01	0.01
Transformer (T+A: Att. fusion)	**0.76 (0.0180)**	0.58 (0.0140)	0.47 (0.0090)	0.57	**0.61**	**0.55**	0.16	**0.18**	0.12
M3ANet (Ours)	0.77 (0.0180)	**0.57 (0.0160)**	**0.46 (0.0110)**	**0.59**	0.58	0.50	**0.18**	0.17	**0.13**

Table 2: Effect of multimodality and multi-speaker inputs (mean and stdev. of 5 runs for each approach).

days allowing experimentation across both short and medium-term periods.

Similar to prior work (Sawhney et al., 2020d; Theil et al., 2019; Yang et al., 2020), we evaluate predicted volatility using the mean squared error (MSE) for each hold period, $n \in \{3,7,15\}$. For the classification task, we report the F1 score and Mathew's Correlation Coefficient (MCC) for the classification task (Matthews, 1975). We use MCC because, unlike the F1 score, MCC avoids bias due to any data skew that may be present as it does not depend on the choice of the positive class. For a given confusion matrix $\begin{pmatrix} tp & fn \\ fp & tn \end{pmatrix}$:

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}} \quad (11)$$

7 Results and Analysis

7.1 Performance Comparison

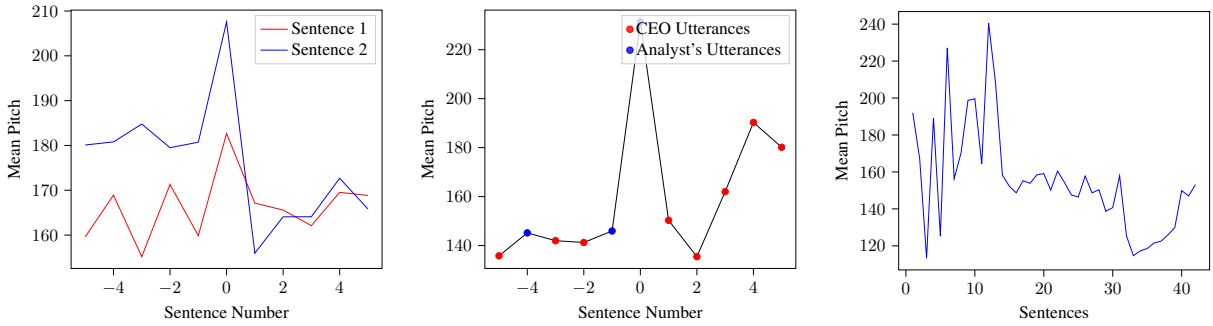
As shown in Table 1, M3ANet achieves the best performance for both the volatility prediction and the price prediction task. We observe improvements using M3ANet (Table 2) that leverages the text and audio modalities along with speaker information. This improvement can be attributed to attention to emphasize the importance of each modality throughout the series of utterances. It can also be observed that the improvements our architecture results in are not quite large in magnitude. We attribute this to the difficulty that the task inherently possesses. Further research in more sophisticated models may result in greater improvements in the performance on M3A.

7.2 Multimodal and Multi-Speaker Learning

From Table 1 and Table 2, we see that in both the MDRM and Transformer models, the multimodal models performed much better than the unimodal counterparts. This performance improvement follows from previous research (Qin and Yang, 2019) with respect to volatility prediction. Similar observations validate our hypothesis that audio cues provide additional information that helps make a better prediction. It is also apparent from Table 2 that adding speaker context improves the prediction result consistently. Thus, we infer that speaker information does play an essential part in forecasting and adds to the data's richness.

7.3 Ablation Study: Fusion

We experiment with fusion by concatenation and fusion by attention for the Transformer and find the latter performing better in most cases (Table 2). We believe this happens because simple fusion techniques cannot produce features that effectively capture the individual modalities' importance. However, attention fusion uses weights for both the modalities, learned by the architecture, to determine the importance of each modality with respect to its counterpart. Using these weights to perform a weighted addition gives a much better representation of both the modalities and their particular importance in a fused vector.



(a) QA1: The CEO answers a question about the company's competitors. Sentence 2: The CEO invites questions.

(b) QA2. The analyst has a spike in their mean audio pitch while the CEO's mean audio pitch is stable.

(c) QA3. The mean audio pitch of the audio clips.

Figure 5: Qualitative Analysis

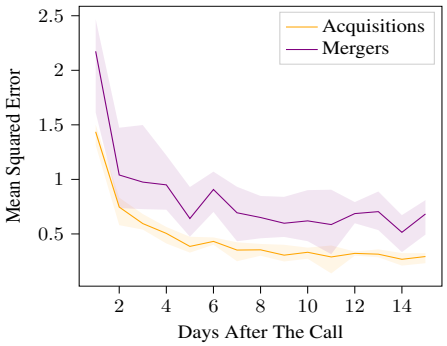


Figure 6: Drift in Predicted Stock Volatility over Time; The line graph represents the mean MSE while the shaded regions represent the performance over 10 runs

Trained On	Tested on Acquisitions Only			Tested on Mergers Only		
	MSE_3	$F1_3$	MCC_3	MSE_3	$F1_3$	MCC_3
Acquisitions	0.65	0.66	0.12	1.47	0.56	0.015
Mergers	0.85	0.28	0.03	1.01	0.47	0.20

Table 3: Ablation Study: Performance of M3A, when trained on Acquisitions and Mergers separately

7.4 Performance Drift over Time

As observed in previous works (Sawhney et al., 2020d) using earnings calls, Figure 6 shows that short-term stock volatility prediction is more complex, possibly due to the erratic price fluctuations after a M&A call. We hypothesize that these price fluctuations settle as more time elapses, similar to the phenomenon of PEAD (Post Earnings Announcement Drift) (Bernard and Thomas, 1989; Bhushan, 1994; Sadka, 2006). This saturation in performance improvement can be attributed to the dilution of cues extracted from the calls, as we 'drift' away from them. However, it can be noted that a similar trend may not necessarily be true for price movement prediction.

7.5 Merger & Acquisition Transfer

We experiment by training `M3ANet` on Mergers and Acquisitions calls separately, and testing both models on each set of calls separately. From Table 3, it can be observed that both models predict the price movement better for their respective sets as expected. It is surprising to see that the models predict volatility of Acquisition calls relatively better than that of Merger calls. This suggests that Acquisition conference calls lead to a volatility that's relatively easier to predict and seems to be an avenue for further research.

7.6 Qualitative Analysis

Call 1: Acquisition of Shape Security by F5 Networks Inc Following the call, F5 Networks Inc suffered a price drop of up to 5.2% within the next month. Studying the call's vocal cues, we notice (Figure 5a) the CEO had sudden peaks in the mean pitch of his audio while answering questions. Similar peaks occurred when a participant asks the CEO about their fraud protection when compared to their competitors. Prior research on audio analysis (Jiang and Pell, 2017) proves a high mean pitch may indicate a lack of confidence in the speaker. It was later ascertained that F5 had overpaid to acquire Shape Security without proper due diligence of fraud protection plans sold by Shape Security. We observe how `M3ANet` successfully predicts the decrease in price for all choices of τ while the unimodal models fail to do the same each time. Though the text reveals no lack of confidence, the audio cues likely allow the model to make a successful prediction.

Call 2: Merger of AK Steel Holding Corporation and Cleveland-Cliffs Inc Following the

merger call, Cleveland-Cliffs Inc saw an increase in their stock price up to 17.9% in the next five days. Similar to the first call, we notice spikes and sudden increases in the audios' mean pitch from Figure 5b. However, the difference exists in the fact that these high pitch patterns come from an analyst in the call and not someone holding an influential position in the companies involved. M3ANet can differentiate between the speakers and correctly predicts the price going up, unlike the transformer variant without speaker embeddings. This shows how the augmentation of the multimodal data with the speaker embedding likely benefits the predictive power of M3ANet.

Call 3: Acquisition of Plateau Excavation Inc by Sterling Construction Company Inc We now analyze this acquisition as an error analysis where M3ANet predicts incorrectly. We see the text transformer performing well on this example and accurately predicting the increase in the stock price for Sterling Construction Company Inc. On the other hand, our multimodal multi-speaker is unable to do the same. Observing the audio cues (Figure 5c), we find a great deal of variance in the mean audio pitch. We attribute the erroneous performance to the potential overfitting of the model or noise in the audio cues.

8 Conclusion

We present a dataset of M&A calls that can be utilized to predict financial risk following M&A calls. We also present a strong baseline model using multimodal multi-speaker inputs from the M&A calls to perform financial forecasting. M3ANet uses attention-based fusion to leverage the interdependency between the verbal message and the vocal cues. Further, the approach uses speaker information to enrich the input data to determine if the speakers' vocal cues or verbal messages conflict with others and accounts for the same. Experiments on M3A display the effectiveness of M3ANet. We hope our M3A can enable more academic progress in the field of financial forecasting.

Ethical Considerations and Limitations

Examining a speaker's tone and speech in conference calls is a well-studied task in past literature (Qin and Yang, 2019; Chariri, 2009). Our work focuses only on calls for which companies publicly release transcripts and audio recordings. The data used in our study corresponds to M&A conference calls of companies in the NASDAQ stock exchange. We acknowledge the presence of gender bias in our study, given the imbalance in the gender ratio of speakers of the calls. We also acknowledge the demographic bias (Sawhney et al., 2021a) in our study as the companies are organizations within the public stock market of United States of America and may not generalize directly to non-native speakers.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *ArXiv*, abs/1908.10063.

Pascal Belin, Bibi Boehme, and Phil McAleer. 2017. The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLOS ONE*, 12:e0185651.

Victor L. Bernard and Jacob K. Thomas. 1989. Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*, 27:1–36.

Ravi Bhushan. 1994. An informational efficiency perspective on the post-earnings announcement drift. *Journal of Accounting and Economics*, 18(1):45 – 65.

E. Biswas, M. E. Karabulut, L. Pollock, and K. Vijay-Shanker. 2020. Achieving reliable sentiment analysis in the software engineering domain using bert. In *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 162–173.

Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2010. Twitter mood predicts the stock market. *Journal of Computational Science*, 2.

Tim Bollerslev. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 52:5–59.

Robert Bowen, Angela Davis, and Dawn Matsumoto. 2001. Do conference calls affect analyst forecasts. *The Accounting Review*, 77.

Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Zhengqi Wen. 2015. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, AVEC '15, page 65–72, New York, NY, USA. Association for Computing Machinery.

Anis Chariri. 2009. Ethical culture and financial reporting: Understanding financial reporting practice within javanese perspective*. *Issues In Social And Environmental Accounting*, 3.

Sudipto Dasgupta, Jarrad Harford, Fangyuan Ma, Daisy Wang, and Haojun Xie. 2020. Mergers under the microscope: Analysing conference call transcripts. *Available at SSRN 3528016*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiao Ding, Yue Zhang, T. Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *IJCAI*.

Bernard Dumas, Alexander Kurshev, and Raman Uppal. 2009. Equilibrium portfolio strategies in the presence of sentiment risk and excess volatility. *The Journal of Finance*, 64:579 – 629.

Robert Engle. 1981. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50.

F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong. 2016. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

R. Fraunhoffer, H. Kim, and D. Schiereck. 2018. Value creation in ma transactions, conference calls, and shareholder protection. *International Journal of Financial Studies*, 6:1–21.

Steven Heston. 1993. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6:327–43.

Jessen Hobson, William Mayew, and Mohan Venkatachalam. 2011. Analyzing speech to detect financial misreporting. *Journal of Accounting Research*, 50.

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R. Hershey, Tim K. Marks, and Kazuhiko Sumi. 2017. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Wenyao Hu, Thomas Shohfi, and Runzu Wang. 2018. What's really in a deal? evidence from textual analysis. *SSRN Electronic Journal*.

Xiaoming Jiang and Marc D. Pell. 2017. The sound of confidence and doubt. *Speech Communication*, 88:106 – 126.

Herb Johnson and David Shanno. 1987. Option pricing when the variance is changing. *Journal of Financial and Quantitative Analysis*, 22:143–151.

Raehyun Kim, Chan Ho So, Minbyul Jeong, Sanghoon Lee, Jinkyu Kim, and Jaewoo Kang. 2019. Hats: A hierarchical graph attention network for stock movement prediction.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280, Boulder, Colorado. Association for Computational Linguistics.

Werner Kristjanpoller, Anton Fadic, and Marcel Minutolo. 2014. Volatility forecast using hybrid neural network models. *Expert Systems with Applications*, 41:2437–2442.

Qingbiao Li, Chunhua Wu, Zhe Wang, and Kangfeng Zheng. 2020. Hierarchical transformer network for utterance-level emotion recognition. *Applied Sciences*, 10:4447.

Jiexi Liu and Songcan Chen. 2019. *Non-stationary Multivariate Time Series Prediction with Selective Recurrent Neural Networks*, pages 636–649.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Rui Luo, Weinan Zhang, Xiaojun Xu, and Jun Wang. 2017. A neural stochastic volatility model.

Burton Malkiel. 2003. The efficient market hypothesis and its critics. *Journal of Economic Perspectives*, 17:59–82.

B.W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442 – 451.

Marc-andre Mittermayer and G.F. Knolmayer. 2007. Newscats: A news categorization and trading system. pages 1002 – 1007.

Sara Moeller, Frederik Schlingemann, and Rene Stulz. 2003. Do shareholders of acquiring firms gain from acquisitions? *SSRN Electronic Journal*.

Mahla Nikou, Gholamreza Mansourfar, and J. Bagherzadeh. 2019. Stock price prediction using deep learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management*, 26.

Nuno Oliveira, P. Cortez, and Nelson Areal. 2017. The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Syst. Appl.*, 73:125–144.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Yu Qin and Yi Yang. 2019. 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*, pages 390–401, Florence, Italy. Association for Computational Linguistics.

Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Alexander Dür, Linda Andersson, and Allan Hanbury. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based IR models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1712–1721, Vancouver, Canada. Association for Computational Linguistics.

Ronnie Sadka. 2006. Momentum and post-earnings-announcement drift anomalies: The role of liquidity risk. *Journal of Financial Economics*, 80(2):309 – 349.

H. Sakoe and S. Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. 2020a. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8426, Online. Association for Computational Linguistics.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. 2020b. Spatiotemporal hypergraph convolution network for stock movement forecasting. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 482–491.

Ramit Sawhney, Arshiya Aggarwal, and Rajiv Ratn Shah. 2021a. An empirical investigation of bias in the multimodal analysis of financial earnings calls. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3751–3757, Online. Association for Computational Linguistics.

Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Ratn Shah. 2020c. VolTAGE: Volatility forecasting via text audio fusion with graph convolution networks for earnings calls. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8001–8013, Online. Association for Computational Linguistics.

Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, R. Shah, and Roger Zimmermann. 2020d. Multimodal multi-task financial risk forecasting. *Proceedings of the 28th ACM International Conference on Multimedia*.

Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Ratn Shah. 2021b. Quantitative day trading from natural language using reinforcement learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4018–4030, Online. Association for Computational Linguistics.

Louis Scott. 1987. Option pricing when the variance changes randomly: Theory, estimation, and an application. *Journal of Financial and Quantitative Analysis*, 22:419–438.

Siegfried Sporer and Barbara Schwandt. 2006. Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology*, 20:421 – 446.

Kilian Theil, Samuel Broscheit, and H. Stuckenschmidt. 2019. Profet: Predicting the risk of firms from event transcripts. In *IJCAI*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Chuan-Ju Wang, Ming-Feng Tsai, Tse Liu, and Chin-Ting Chang. 2013. Financial sentiment analysis for risk prediction.

Lee Wurm, Douglas Vakoch, Maureen Strasser, Robert Calin-Jageman, and Shannon Ross. 2010. Speech perception and vocal expression of emotion. *Cognition Emotion*, 15:831–852.

Yumo Xu and Shay B Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979.

Linyi Yang, Tin Lok James Ng, Barry Smyth, and Ruihai Dong. 2020. Html: Hierarchical transformer-based multi-task learning for volatility prediction. *Proceedings of The Web Conference 2020*.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. pages 5415–5421.

Jie Zheng, Andi Xia, Lin Shao, Tao Wan, and Zengchang Qin. 2019. Stock volatility prediction based on self-attention networks with social information. pages 1–7.