MAEC: A Multimodal Aligned Earnings Conference Call Dataset for Financial Risk Prediction

Jiazheng Li*
University College Dublin
Dublin, Ireland
jiazheng.li@ucdconnect.ie

Linyi Yang*†
Insight Centre for Data Analytics
University College Dublin
Dublin, Ireland
linyi.yang@insight-centre.org

Barry Smyth
Insight Centre for Data Analytics
University College Dublin
Dublin, Ireland
barry.smyth@insight-centre.org

Ruihai Dong
Insight Centre for Data Analytics
University College Dublin
Dublin, Ireland
ruihai.dong@insight-centre.org

ABSTRACT

In the area of natural language processing, various financial datasets have informed recent research and analysis including financial news, financial reports, social media, and audio data from earnings calls. We introduce a new, large-scale multi-modal, text-audio paired, earnings-call dataset named MAEC, based on S&P 1500 companies. We describe the main features of MAEC, how it was collected and assembled, paying particular attention to the text-audio alignment process used. We present the approach used in this work as providing a suitable framework for processing similar forms of data in the future. The resulting dataset is more than six times larger than those currently available to the research community and we discuss its potential in terms of current and future research challenges and opportunities. All resources of this work are available at https://github.com/Earnings-Call-Dataset/.

CCS CONCEPTS

• **Computing methodologies** → **Language resources**; • **Information systems** → **Multimedia content creation**.

KEYWORDS

Multimodal Aligned Datasets; Earnings Conference Calls; Financial Risk Prediction

ACM Reference Format:
Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020. MAEC: A Multimodal Aligned Earnings Conference Call Dataset for Financial Risk Prediction. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3340531.3412879

*Both authors contributed equally to the paper
†Corresponding Author

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412879

1 INTRODUCTION

In recent years natural language processing and analysis techniques have transformed computational research in a variety of domains from business and commerce to health and medicine. This has been enabled for at least two important reasons: (1) the development of new algorithms and techniques that are capable of analysing natural language at scale [1, 9, 27]; and (2) the availability of large-scale real-world datasets [8, 25]. Recently multimedia datasets have proven to be especially important and text, images, audio, video datasets have been combined in a variety of ways to address many challenging tasks in a variety of domains [17, 23].

In this work, we focus on the world of financial data with stock volatility prediction as an important and challenging problem of interest. In particular, we propose a novel dataset that is well suited to this task, by combining text and audio from the earnings calls of S&P 1500 companies. Our current research is focused on using the text and audio features from these calls to predict how a company's stock will behave in the days and weeks that follow. The dataset presented here has been collected over the past 40 months and represents a significant extension to the related dataset used by [23]. It includes data from 3,443 earnings calls, made up of cleaned transcription texts and aligned, sentence-level audio slices, plus important audio features.

Earnings conference calls are always given by leaders or senior managers in the listed companies, typically the CEO or CFO. They usually begin with a brief presentation to describe company performance of the last quarter and expectations for the coming period and following with a Q&A session. Previous work [19] found that the question-answer portion of the earnings call is more informative in terms of the correlation with intra-day absolute returns as the Q&A parts are not well prepared by executives. In this session, the senior manager may provide crucial insight into the company that is not available from the presentation. As shown in Figure 1, the senior manager's response is remixed with informal discourse words and the use of euphemisms.

Financial Analyst: So the first one, on **Prime**, **let** me ask you this. As you think about your **U.S. Prime penetration**, there's some data that shows you're doing a very good job at capturing a lot of the **middle to higher** income households and **now** you're raising price. Talk about the tension point you need to solve to sort of reach some of the lower income households and even households that are not yet **Prime**. What are the main reasons why people in the U.S. are not signing on for **Prime** at this point? And the second one, on **early** learnings from the integration of **Prime Now** and **Whole Foods**, recognizing it's **only** in a few cities. What can you share about what you're seeing about purchase behavior, **early** learnings? And what are the main signposts you're watching as you determine how quickly to roll that out to more cities in the **U.S.**, and **hopefully**, **New York** soon?

Amazon CFO: Sure. On your first question about **Prime** penetration, without getting into any statistics on penetration and by country, **I would say** we do have other options for, if **you'll** notice, there's the monthly option, **obviously** provides more flexibility for people who want to try out Prime before committing to the annual plan. There's discounted student plans. There's also discounts for other groups. So **we do feel** it's still the best deal in retail, and we **just** work to make it better and better each day. The second thing you mentioned is a good example. So the ability in 10 cities to get **Prime Now** deliveries of **Whole Foods** groceries is an added benefit for people in that market using Prime – those markets using Prime Now. So as far as the **Whole Foods**, specifically on the question of what will – what **we'll** look at as far as expanding that grocery delivery, **we're going to** use the 10 cities as a test and see how customers respond, **just** like we always do, and make sure that our deliveries are great for those people, and then **we'll** announce expansion plans **once** we digest that, the feedback we get from customers.

Figure 1: Earnings calls are naturally given and impromptu. In this example Q&A session from a Amazon earnings call on April 26, 2018, the analyst asks three questions. We can notice the speech dis-fluency and euphemism. Their conversion mainly focus on specific object or tendency based on timing.

We believe that our data release will help advance research at the intersection of language, audio, and finance. The sentence-level pairwise alignment between the text and audio could be useful in pragmatics, semantics, and acoustic analysis of financial documents. Also, it can be used to train deep learning models that forecast the financial risk given the verbal and vocal cues. The technical contributions mainly focus on creating a multimodal aligned earnings call dataset (MAEC) and providing a suitable framework for processing similar forms of data in the future.

The remainder of this paper is organized as follows. We discuss the relevant models and datasets compared to us in Section 2. A structured pipeline and any challenges that may come out during data processing are explained in Section 3. A brief data analysis example is provided in Section 4 to show interesting details of our dataset. We applied our dataset in the task of volatility prediction and discuss the results in Section 5. We conclude our work at the end.

2 RELATED WORK

We view the earnings conference call analysis as a multimodal analysis problem incorporating textual and audio information. Accordingly we review related work in text-based and multi-modal analyses.

2.1 Text-based Financial Datasets

Most financial disclosures are textual materials, and the complexity of textual financial data in recent years makes it increasingly challenging for investment analysts to extract valuable insights and perform analysis without the high quality processed datasets. For instance, in recent studies of 10-K reports and earnings call transcripts for risk prediction, the information from financial disclosures dataset supports many interesting findings [14, 16] as follows.

Previous work [14] points out that the pragmatics and semantics of earnings calls have a significant influence on analysts' decision-making behavior. Analysts reprice target recommendations based on the company's advice after the quarterly earnings calls. In fact, in a very early study, [16] mentions that language-based models can reveal deceptive information during earnings calls and cause the stock price swings in financial markets later. However, unlike other text-driven tasks that have rich datasets to use, for instance, financial news datasets proposed by [11, 12, 30], tweets data used by [5, 21], and 10-K report datasets published by [25]. In the emerging research task of using earnings call to predict financial risk, a large-scale online available dataset is rarely found.

2.2 Multi-modal Financial Datasets

Recently, progress in multi-media processing has been based on incorporating different types of features together during the training. Some of the popular datasets have been created to adapt to the progress of multimodal learning and high-level embedding from various data sources. For example, the Vision-and-Language BERT (ViBERT) [26] requires images and natural language datasets, and M-BERT [24] explores the possibility of leveraging the language, acoustic and visual data together based on the transformer architecture. Our work focuses on the creation of a text-audio aligned dataset, motivated by S&P 500 Earnings Conference Calls dataset [23] limited in the year of 2017, which was the first work to view earnings call analysis as a multimodal (text + audio) opportunity.

The previous work [2] has proven that different emotions and psychological activities [3, 13] of a speaker can be abstracted and represented. Meanwhile, the recent models proposed by [23, 28] shed new light on using text-aligned data to capture not only the verbal features but also the vocal cues. In addition, [14] leverages the semantic and pragmatic features extracted from the transcripts to provide the interesting findings that earnings calls are moderately predictive of analysts' decisions. To the best of our knowledge,

Table 1: Comparisons of our multimodal aligned earnings calls dataset and the existing public earnings calls datasets

Dataset	MAEC	Qin's[23]	Keith's[14]
Duration	2015-2018	2017	2010-2017
#Companies	1,213	280	642
#Instances	3,443	576	12,285
Transcripts	√	√	√
Audios	√	√	
Basic Audio Features	√	√	
MFCC Audio Features	√		

Table 2: Statistic of the MAEC

#Multimodal Aligned Earnings Calls	3,443
Start Date-	25th Feb 2015
End Date-	21th Jun 2018
#Companies	1,213
#Sentences	394,277
#Tokens	8,019,142
Total Length of the Audio	920.67 Hours

our dataset is the current largest open dataset in the task of the text-audio aligned earnings calls. As shown in Table 1, we collect a multimodal aligned dataset earnings call of up to three years into consideration whereas [23] takes only one year and [14] only considers the textual data.

We believe that these recent works provide a starting point for the task of financial risk forecasting using the multimodal dataset so that a large-scale multimodal aligned dataset of earnings call created in this work is a crucial and urgent need for this research community.

3 DATA PROCESSING

We describe our approach in detail for collecting large-scale text and audio recordings in Figure 2. Our goal is to align the sentence-level transcripts and audio clippings for the presentation and Q&A part of an earnings call. The task is challenging, as different earnings calls vary in their meeting agendas, and also various executives have different phonetic characteristics. Meanwhile, the audio recordings can be noisy, with some sentences having poor sound quality. Moreover, the vast majority of earnings calls have very long transcripts and the corresponding audio recordings. As a result, the process of alignment becomes very time-consuming. Hence, most previous alignment approaches have [23] dealt with small datasets.

3.1 Data Acquisition

Our S&P 1500 earnings conference call dataset is sourced as follows.

Earnings Call Transcripts: We downloaded the earnings call transcripts from the website Seeking Alpha.[1] The transcripts labeled company and conference id, speech structure, the role of the speaker, and the content of the speech. We have received the written consent from the Seeking Alpha.

Earnings Call Audio: We downloaded the earnings call audio from the website EarningsCast.[2] The audio is a recording of each conference call without any labels.

A total of 9,383 collected pairs of transcripts and audio recordings have formed our source data. Both data sources are free to download and available to the public.

3.2 Transcript Processing

The transcript records the presentation delivered by executives and the Q&A interaction between executives and financial analysts or reporters chronologically. The transcripts collecting from Seeking Alpha includes two types of files: content files, which contain all of the conversational content from earnings calls and speech sequence files, which include the speaker's duty for each sentence in the presentation part and Q&A session. Furthermore, to apply best practices in keeping with the EU spirit of data protection, we mask all popular human names by replacing it with the 'UNK' tokens in our published transcripts. Volatility in the company's stock returns is often caused by earnings that do not meet market expectations or spokesman fails to address critical questions during the Q&A phase. Hence, following [23], we retain and process the data of the spokesman's presentation and their responses in Q&A, which is known to be the most crucial part [19].

With the label of the speaker of each paragraph, we can easily prune unnecessary speeches made by investigators or conference operators when we were doing the paragraph-level audio alignment. In the end, the transcript text file has been cleaned and split to each sentence during the sentence-level and text-audio alignment described in Section 3.4.

3.3 Paragraph Level Text-Audio Alignment

During audio processing, document-level transcript text files were processed in parallel to ensure that the text was paired with the audio. The speech sequence file helps the first stage of text-audio alignment. We leverage the algorithm used on S&P 500 Earning Conference Call dataset by [23]. In an audio recording of the earnings conference call, the start and the end are usually mixed up with short sentences spoken by multiple people. The good solution for a more accurate audio segment is Iterative Forced Alignment [20]. We used Aeneas [3] to process the segmentation. As shown in Algorithm 1, we show a detailed explanation of the algorithm we used in paragraph-level segmentation.

The input source for our algorithm is the pre-processed transcript texts, processed transcript speech sequences, and the corresponding audios. The algorithm traverses through all folders and produces a synchronized map. We use a speech sequence file to identify the speaker's position, where the algorithm slices every sentence and audio segment. In this step, each text remains at a paragraph-level but with a non-management person removed. Meanwhile, audio is sliced corresponding to each paragraph and saving only slices spoken by management person.

[1] Seeking Alpha: https://seekingalpha.com/

[2] EarningsCast: https://earningscast.com/

[3] Aeneas: https://github.com/readbeyond/aeneas

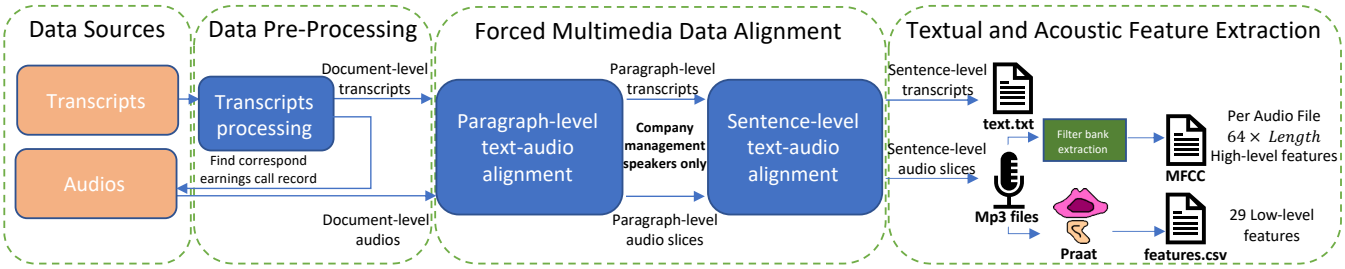

Figure 2: The pipeline of the data processing

Algorithm 1 Paragraph-level segmentation

function ALIGNMENT(a_i, t_i, s_i)
  $syncMap \leftarrow Aeneas(a_i, t_i)$
  $speaker \leftarrow GetLastSpeaker(s_i)$
  $slice \leftarrow GetLastParagragh(t_i)$
  if *speaker* is in management team **then**
    $SaveAudioSlice(a_i, syncMap.start, syncMap.end)$
    $SaveTextSlice(slice)$
  end if
  $RemoveLastSpeaker(s_i)$
  $RemoveLastParagragh(a_i, t_i)$
  if $syncMap.start < 1s$ && Not last paragraph **then**
    Prune current alignment
  else if $syncMap.start < 1s$ && Is last paragraph **then**
    return *True*
  else
    return *False*
  end if
end function
function ITERATIVE SEGMENTATION
  for *i* in all text-audio pairs **do**
    $a_i, t_i \leftarrow Audio_i, Transcript_i$
    $s_i \leftarrow SpeechSequence_i$
    while $AlignmentResult! = True$ **do**
      $AlignmentResult \leftarrow Alignment(a_i, t_i, s_i)$
    end while
  end for
end function

The audio is sliced iteratively from the end to the beginning. We use one second to evaluate the finish of alignment. The segmentation process that occurs is defined as a failure if the start point is less than one second, and not aligned text exists. There are a few causes that bring about this failure. Such examples are audio downloads not pairing with the transcripts or the audio broken up in places with several minutes of ringing noise or silence. In these cases, Aeneas is unable to produce faultless maps for audio and text. It will result in a minuscule slice or a sizeable paragraph with only a one-second duration. This step of our algorithm automatically pruned most of these cases. Also, when the audio is shorter than one second, we believe it is impossible to carry any meaningful information for an audio feature and, it is impossible to understand by a human.

In the end, we produced 9,225 out of a total of 9,383 collections of original files. Most of the pruned pairs are due to bad audio quality or incorrect pairings of text and audio.

3.4 Sentence Level Text-audio Alignment

The paragraph-level algorithm produces textual data and audio slices for the sentence-level alignment. Since we finished identifying speaker and selection, transcripts speech sequence files are no longer needed to be involved in further processes.

The structure of the sentence-level text-audio segmentation algorithm is similar to paragraph-level alignment. We apply Iterative Forced Alignment [20] on further processes. Since we already cleaned our text and filtered non-management speakers' speeches in our text, we removed speaker determination procedures in this algorithm and procedures on transcripts speech sequence files. Aeneas has been used to produce a synchronous map with input text pieces and input audio slices as well.

Unlike paragraph-level segmentation, we create a strict condition to check that the period between the start and the end should be at least 0.7 second. By investigating each audio extracts, we found that most shorter phrases like "`Thank You`" or "`Morning`" are spoken with a period of longer than 0.7 second. We treat all audio less than 0.7 second as a failed alignment and prune that text-audio collection. Based on experience, most audio slices that are less than half a second occur due to Aeneas incorrectly pairing of sentences and audio. Similar to the reason explained in Section 3.3, speakers' accents, pacing, or remix of conversation may result in the inaccurate start and end duration in a synchronous map. As this is an iterative forced alignment, such an inaccuracy will result in an incorrect text and audio alignment at the beginning of the audio file. Therefore, we subject such values to a hard check and exclude all inaccuracies.

The goal is to produce as many instances as possible. Not included are any comparison of iterative forced alignment with results from other alignment methods to filter less accurate alignment slices. The previous experience shows that small alignment differences will not cause a significant effect on the result of a further experiment on this dataset. Finally, 3,443 multimodel-aligned, sentence-level transcripts and audio collections were produced in total.

3.5 Challenges

We describe the main challenges during the data construction from three aspects.

3.5.1 Inaccurate Alignment. There are a few main challenges we met when we are producing this dataset. Firstly, audio quality results in inaccurate alignment processing. There is no standard regulation for releasing conference call audios. Some of the audio files from small companies are assorted with significant noise. This noise would not influence human understanding, but it can affect extraction as the audio would be considered inaccurate. We deal with this challenge by restricting our algorithm, but, we believe future extension can focus on noise reduction for better production.

3.5.2 Multiple Speakers. Multiple speakers mixing also hamper to audio alignment. The Aeneas package we used compares sound between the computer produced and the audio file. The remixed speech will result in incomparable waves. Since people speak differently, a deep-learning based method could identify the person and clean and extract audio from them. We take it as our future work to further refine our alignment algorithm.

3.5.3 Memory Management. To produce such a large-scale text-audio aligned dataset is time-consuming work, we divide the whole processing task into small tasks by using Spark [29]. In particular, it costs two months to construct paired data using a paragraph-level text-audio alignment algorithm and another one month on sentence-level alignment. We expect that the distributed computing technique with enough computing resources could speed up the process of data construction in similar tasks.

3.6 Ethics

We will discuss the potential risk of our dataset and the best practice we have applied to protect the data privacy in this section. There is some personal information (names and positions) that could be found in our data, as the original earnings conference calls of the public companies in the United States are required to be publicly available by the U.S. law. We, therefore, argue that the personal information in the earnings call dataset relates to legal persons and not natural persons. Despite this, we still have applied our best practice to protect the personal information following the European spirit of the General Data Protection Regulation (GDPR). In particular, we replaced all of the identify-able names with the 'UNK' tokens when processing the transcripts of the earnings calls. Publishing under the Creative Commons Attribution-ShareAlike 4.0 International License is the anonymous data resource. [4]

3.7 Feature Extraction

Our feature extraction is of two parts, text and audio. The features of the textual data in earnings call can be encoded by the large-scale pre-trained model, like BERT [10].

We assume the primary function of our dataset is in deep learning. A proven fact is that speech analytics is possible to use on phonetic recognition [23]. In this step, we used Praat [4] as our tool to extract low-level audio features to enable model training. A total of 29 extracted features included pitch, intensity, jitter, HNR(Harmonic to Noise Ratio), and much more from sentence-level audio slices. We also extricated filter bank features with Python speech feature packages. Concerning high-level filter bank features, they are more effective as sliced audio from them can be represented in graphs. We believe our dataset can enable researchers to apply traditional natural language processing tasks. Both of our low-level audio features and filter bank features are uploaded and open to the research community.

[4]MAEC Dataset: https://github.com/Earnings-Call-Dataset/MAEC-A-Multimodal-Aligned-Earnings-Conference-Call-Dataset-for-Financial-Risk-Prediction

4 DATA ANALYSIS

We briefly describe an analysis of the dataset produced. A correlation analysis between the semantic features with the stock returns is also introduced.

4.1 Transcripts

Based on the textual analysis with conference call transcripts from 2007 to 2018 (larger than multimodal data), we found some interesting properties. From an economic perspective, we notice that the word "growth" always being ranked at the top 10 most frequent words mentioned in the transcripts of the earnings call to show a good market outlook. Meanwhile, we also observe that the phrases describing future expectations such as "long term" and "forward looking" increase in occurrence during the years of 2009-2011. It is also interesting to find that uncertainty words – such as "might", "maybe" and "probably" – increase in occurrence during the global economic crisis.

Besides, as shown in Figure 3, we count the occurrences of emotional words for each year, and further, calculate the yearly change of the ratio $\Delta(\frac{number of positive words}{number of negative words})$ to explore the correlation between the semantic features in yearly earnings calls with the stock returns.

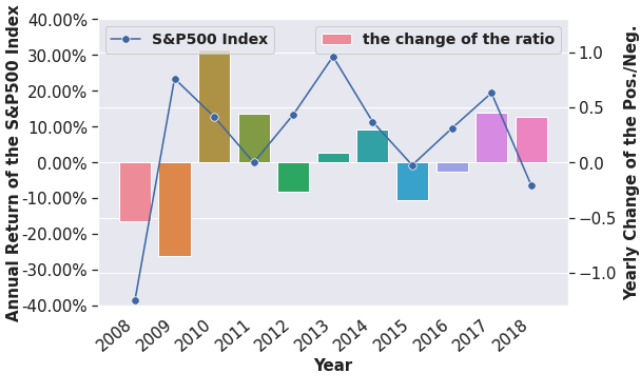

Figure 3: The annual return of the S&P500 index with the change of the semantic features in earnings call

4.2 Audio Recordings

We also analyzed the dataset of the audio recordings as shown in Table 3. Noise in the vocal data may impact the accuracy of the resulting paired data after alignment. The very short audio slices (less than one second) mainly contain meaningless modal particles like "yes", "morning". Meanwhile, such an inaccurate map may result in a few seconds distinction between the real duration in alignment processes. Also, in the step of the feature extraction, we may notice that some audio clips cannot generate audio features

Table 3: Data Analysis of S&P1500 Earnings Call

Quarter	Number of Calls	Average Sentences per Call	Cumulative Percentage	Total Audio Length	Audio Length Per Sentence
2015Q1	38	89	1.10%	7h	7.92s
2015Q2	238	96	8.02%	51h	8.01s
2015Q3	215	92	14.26%	45h	8.18s
2015Q4	274	96	22.22%	57h	7.78s
2016Q1	296	95	30.82%	64h	8.11s
2016Q2	452	105	43.94%	108h	8.18s
2016Q3	438	98	56.67%	101h	8.37s
2016Q4	214	83	62.88%	39h	7.83s
2017Q1	371	91	73.66%	82h	8.62s
2017Q2	274	163	81.61%	110h	8.82s
2017Q3	157	160	86.17%	60h	8.62s
2017Q4	156	160	90.71%	61h	8.72s
2018Q1	132	165	94.54%	54h	8.98s
2018Q2	188	180	100.00%	82h	8.72s

(MFCC or the low-level features) successfully because the length of some audio clippings is too short. Hence, there is a hint for users that they may filter the dataset by the length of the audio clips.

4.3 Stock Returns Correlation

We are interested in whether and how the sentiment of the earnings call can influence the expectations and judgments of the market – namely S&P500 Index and S&P1500 Index. We calculate the ratio of positive sentiment terms and negative sentiment terms to the number of sentiment words in the company executives' presentations and responses. Then, the change of the ratio compared to last year is used to represent the current year's market sentiment. The financial lexicons generated by [18] from fourteen years of historical 10-K reports is leveraged to select the emotional words.

The results of the semantic correlation analysis are given in Table 4. We show the correlation between the annual return of the S&P 500(1500) index and the yearly change of the market sentiment in earnings call respectively. There are several interesting findings to discuss. First, we find that both S&P 500 and S&P 1500 index in the current year's annual return has a somewhat positive correlation with the market sentiment found in this year's earnings calls. Meanwhile, it is noteworthy that the market sentiment in the current year has strong negative correlations to the next year's financial indexes, namely, -0.4495 to S&P500 index and -0.4877 to S&P 1500 index. This provides strong evidence in support of the idea that underlying semantic features in earnings calls can improve financial index prediction.

5 EXAMPLE USE CASE

Previous research [23] has demonstrated the benefits of combining text with audio data, compared to text-only features, in volatility prediction. However, the public dataset [23] they used only contains 576 recordings of earnings call during the year of 2017. To further explore the use of audio features in a range of related or complementary tasks (e.g. volatility prediction, asset pricing, stock recommendation, etc.), we extend their dataset by selecting companies from S&P1500 instead of S&P500 and also collect the resource data during from Feb 2015 to Jun 2018.

Table 4: The semantic correlation analysis from 2007 to 2018. The change of the ratio compared to last year is used to represent the current year's market sentiment.

Index Price	Pearson's Correlation with the Semantic
Current Year's S&P500 Index	0.1782
Current Year's S&P1500 Index	0.1726
Next Year's S&P500 Index	-0.4495
Next Year's S&P1500 Index	-0.4877

In this section, we describe a series of benchmarks using our dataset for the task of volatility prediction, including text-based and multimodal approaches. Furthermore, the results based on our dataset will be discussed later and compared to the previous works.

5.1 Task Definition

Following [23, 28], we formulate the volatility forecasting problem as a multivariate regression task, with textual and audio data as raw inputs, and n-day volatility predictions – the predicted average volatility over the following n days – and single-day volatility prediction for day-n as the dual prediction outputs. Different models have been built for predicting the volatility of different years. The year-basis experiment is implemented based on the assumption that the historical experience would be changed rapidly in Fintech.

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

In Equation 1, r_i is the stock return on day i and $\bar{r}$ is the average stock return in a window of n days. The return is defined as $r_i = (P_i - P_{i-1})/P_{i-1}$, where P_i is the adjusted closing price of a stock

Table 5: The year-basis data splitting strategy in chronological order

Year	Training set	Validation set	Testing set
2015	25/02/2015 - 22/10/2015 (535 Instances)	22/10/2015 - 28/10/2015 (76 Instances)	28/10/2015 - 17/12/2015 (154 Instances)
2016	05/01/2016 - 03/08/2016 (980 Instances)	03/08/2016 - 12/08/2016 (140 Instances)	15/08/2016-15/11/2015 (280 Instances)
2017-2018	17/01/2017 - 07/11/2017 (894 Instances)	07/11/2017 - 15/02/2018 (127 Instances)	15/02/2018 - 21/06/2018 (257 Instances)

Table 6: The year-basis experimental results from 2015 to 2018 using Mean Square Error (MSE↓) as the evaluation metric

Year	2015				2016				2017-18			
Price-based Methods	**n=3**	**n=7**	**n=15**	**n=30**	**n=3**	**n=7**	**n=15**	**n=30**	**n=3**	**n=7**	**n=15**	**n=30**
LSTM+ATT	1.599	0.560	0.339	0.284	1.544	0.571	0.362	0.288	1.481	0.562	0.390	0.293
Text-based Methods	**n=3**	**n=7**	**n=15**	**n=30**	**n=3**	**n=7**	**n=15**	**n=30**	**n=3**	**n=7**	**n=15**	**n=30**
CNN-Text [15]	1.427	0.462	0.340	0.266	1.603	0.562	0.370	0.267	1.538	0.600	0.385	0.315
Multi-modal Methods	**n=3**	**n=7**	**n=15**	**n=30**	**n=3**	**n=7**	**n=15**	**n=30**	**n=3**	**n=7**	**n=15**	**n=30**
MDRM–w/o audio	1.438	0.501	0.314	0.298	1.469	0.498	0.354	0.304	1.541	0.536	0.374	0.294
MDRM [23]	1.425	0.488	0.320	0.285	1.426	0.476	0.311	0.259	1.430	0.475	0.323	0.283
HTML–w/o audio	1.199	0.440	0.231	0.187	1.287	0.479	0.300	0.249	1.236	0.510	0.298	0.256
HTML [28]	1.065	0.416	0.272	0.196	1.160	0.515	0.314	0.236	1.152	0.466	0.302	0.250

on day i. In Equation 2, v_n can also be considered a noisy proxy of log volatility [7].

Different from the other methods that only have one task, in the implementation of HTML, we follow the work of [28] to use multi-task learning. More specifically, we use the single day log volatility estimated by the daily log absolute return as our auxiliary task additionally see Equation 2. The objective of multi-task learning of the HTML model is to simultaneously predict these two quantities $v_{[0,n]}$ and v_n.

$$v_n = \ln\left(\left|\frac{P_n - P_{n-1}}{P_{n-1}}\right|\right) \tag{2}$$

5.2 Methodology

The previous research [6] in the finance domain reveals that the predictions of share prices and volatility are made typically using models with thousands of predictors and resulted in a highly sensitive performance. Semantic information in the text and emotional information in the audio can be seen as two of the thousands of effective predictors that may lead to sensitive results. Hence, to avoid the macroeconomic influence so much, we first split our dataset by years except 2018 into chronological order, as the data of only half of the year in 2018 collected is too tiny to use independently. Then, we further split the dataset of each year into mutually exclusive training/validation/testing sets in the ratio 7:1:2, instead of splitting the whole dataset directly. Note that the 7:1:2 split refers to the earning calls. Specifically, we conclude our splitting strategy in Table 5.

We select four representative works feeding with different types of input data respectively to show the multiple uses of our dataset in the task of volatility prediction, the selected methods include LSTM+ATT with historical price data, CNN-Text with textual input, and state-of-the-art multimodal methods [23, 28] with both textual and acoustic data. All of the methods are compared fairly using the Mean Square Error (MSE) as the evaluation metric.

5.3 Results and Discussions

The results of using our dataset to predict the volatility are presented in Table 6, for 3, 7, 15, and 30-day time-periods. It should be clear that we build the year-basis experiment using the different groups of data from 2015-2018. We find that the HTML model achieves the highest prediction performance (lowest MSE values) for each of the target time-periods consistently from 2015 to 2018. In addition to such overall measures of performance, however, we are also interested in discovering more insights from our large-scale data. Thus, in the following subsections, we discuss further discoveries from the experimental results.

5.3.1 The Benefits of Audio Features. We explore the benefits of audio features in multimodal methods for MDRM [23] and HTML [28] respectively. We note a distinct improvement of using the audio features as the additional input using MDRM in almost all settings (excluding in the year of 2015, when n=15). However, if we adopt the hierarchical transformer (HTML), we note that the improvements are not significant, especially for the long-term prediction (n=15, n=30). However, it indeed improves performance for almost all short-term settings from 2015 to 2018. This might be interesting to find that short-term volatility is more greatly influenced by the vocal cues compared to the long-term volatility consistently from 2015 to 2018 with the hierarchical transformer architecture.

5.3.2 The Challenge of Volatility Prediction. As the splitting strategy shown in Table 5, we note that as the years grew, more instances of companies in S&P 1500 were included in our dataset. It leads to the year-basis prediction results generated by all methods except the price-based methods decline in terms of the MSE in almost all of the settings over the years. This may hint that the gaps of different companies between the training and testing set can have a negative influence on the performance. Meanwhile, from the previous studies in economics [22], we know that the volatility of small companies' has gotten more volatile and harder to predict than the big companies. In this experiment, we are using the data collected from the S&P 1500 companies instead of S&P 500 in the previous work [23], it leads to the results that on our new dataset has wider application scenarios for predicting volatility and also face new challenges.

6 CONCLUSION

In the task of financial risk prediction, recent advances in machine learning mean that researcher attention has moved from classical time-series prediction approaches to more sophisticated methods that incorporate multimedia (often unstructured) data such as financial reports, social media, and even audio recordings.

In this paper, we release a multi-modal, aligned, earnings call dataset constructed using the two-stage unsupervised algorithm for aligning sentence-level transcripts and the corresponding audio clips, along with the processed linguistic and acoustic features. The dataset is available online and should assist academic and industry researchers at the intersection of machine learning and financial modelling.

Although our dataset focuses on the financial domain, our approaches should generalize to other domains with abundant volumes of unstructured but align-able multi-modal (text-audio) data, such as speech recognition and lie detection. We also envision extending this work by including more speakers (not only executives) and, by considering the structure of conference calls. Ultimately, we believe this work will benefit the emerging research area of multi-modal analysis in finance where any early insights can be valuable signals when it comes to predicting the future.

ACKNOWLEDGEMENT

This research was supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_2.

REFERENCES

[1] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063* (2019).
[2] Jo-Anne Bachorowski. 1999. Vocal expression and perception of emotion. *Current directions in psychological science* 8, 2 (1999), 53–57.
[3] P Belin, B Boehme, and P McAleer. 2017. The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLoS ONE* 12, 10 (2017), e0185651.
[4] Paul Boersma and Vincent Van Heuven. 2001. Speak and unspeak with praat. *Glot International* 5, 9/10 (2001), 341–347.
[5] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8.
[6] Tim Bollerslev, Andrew J Patton, and Rogier Quaedvlieg. 2016. Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics* 192, 1 (2016), 1–18.
[7] Michael W Brandt and Christopher S Jones. 2006. Volatility Forecasting With Range-Based EGARCH Models. *Journal of Business & Economic Statistics* 24, 4 (2006), 470–486.
[8] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Numeracy-600k: learning numeracy for detecting exaggerated information in market comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL '19)*. 6307–6313.
[9] Yingmei Chen, Zhongyu Wei, and Xuanjing Huang. 2018. Incorporating Corporation Relationship via Graph Convolutional Neural Networks for Stock Price Prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. 1655–1658.
[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[11] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence (IJCAI '15)*.
[12] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to Chaotic Whispers: A Deep Learning Framework for News-Oriented Stock Trend Prediction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 261–269.
[13] Xiaoming Jiang and Marc D Pell. 2017. The sound of confidence and doubt. *Speech Communication* 88 (2017), 106–126.
[14] Katherine Keith and Amanda Stent. 2019. Modeling Financial Analysts' Decision Making via the Pragmatics and Semantics of Earnings Calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL '19)*. Florence, Italy, 493–503.
[15] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP '14)*. 1746–1751.
[16] David F Larcker and Anastasia A Zakolyukina. 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research* 50, 2 (2012), 495–540.
[17] Sang Il Lee and Seong Joon Yoo. 2019. Multimodal deep learning for finance: integrating and forecasting international stock markets. *The Journal of Supercomputing* (2019), 1–19.
[18] Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* 66, 1 (2011), 35–65.
[19] Dawn Matsumoto, Maarten Pronk, and Erik Roelofsen. 2011. What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions. *The Accounting Review* 86, 4 (2011), 1383–1414.
[20] Pedro J Moreno, Chris Joerg, Jean-Manuel Van Thong, and Oren Glickman. 1998. A recursive algorithm for the forced alignment of very long audio segments. In *Fifth International Conference on Spoken Language Processing*.
[21] Nuno Oliveira, Paulo 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 Systems with Applications* 73 (2017), 125–144.
[22] Ser-Huang Poon and Clive WJ Granger. 2003. Forecasting volatility in financial markets: A review. *Journal of economic literature* 41, 2 (2003), 478–539.
[23] 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 (ACL '19)*. Association for Computational Linguistics, Florence, Italy, 390–401.
[24] Wasifur Rahman, Md Kamrul Hasan, Amir Zadeh, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 2019. M-BERT: Injecting Multimodal Information in the BERT Structure. *arXiv preprint arXiv:1908.05787* (2019).
[25] 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)*. 1712–1721.
[26] Quoc-Tuan Truong and Hady Lauw. 2019. Multimodal Review Generation for Recommender Systems. In *The World Wide Web Conference (WWW '19)*. 1864–1874.
[27] Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. 2018. Hybrid Deep Sequential Modeling for Social Text-Driven Stock Prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. 1627–1630.
[28] Linyi Yang, Tin Lok James Ng, Barry Smyth, and Ruihai Dong. 2020. HTML: Hierarchical Transformer-based Multi-task Learning for Volatility Prediction. In *Proceedings of The Web Conference 2020 (WWW '20)*. 441–451.
[29] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. 2010. Spark: Cluster Computing with Working Sets. In *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud '10)*. USENIX Association, USA, 10.
[30] Xi Zhang, Yunjia Zhang, Senzhang Wang, Yuntao Yao, Binxing Fang, and S Yu Philip. 2018. Improving stock market prediction via heterogeneous information fusion. *Knowledge-Based Systems* 143 (2018), 236–247.