Saliency-Aware Interpolative Augmentation for Multimodal Financial Prediction

Samyak Jain[1*], Parth Chhabra[1*], Atula Neerkaje[2*], Puneet Mathur[3], Ramit Sawhney[4,5], Shivam Agarwal[6], Preslav Nakov[5], Sudheer Chava[4], Dinesh Manocha[3]

IIIT Delhi-India[1], UT Austin-United States[2], University of Maryland-United States[3], Georgia Institute of Technology-United States[4], Mohamed bin Zayed University of Artificial Intelligence-UAE[5], University of Illinois Urbana-Champaign-United States[6]

{parth19069, samyak19098}@iiitd.ac.in, atulaj@utexas.edu, {puneetm, dmanocha}@umd.edu, rsawhney31@gatech.edu / {ramit.sawhney, preslav.nakov}@mbzuai.ac.ae, shivamag99@gmail.com, sudheer.chava@scheller.gatech.edu

Abstract

Predicting price variations of financial instruments for risk modeling and stock trading is challenging due to the stochastic nature of the stock market. While recent advancements in the Financial AI realm have expanded the scope of data and methods they use, such as textual and audio cues from financial earnings calls, limitations exist. Most datasets are small, and show domain distribution shifts due to the nature of their source, suggesting the exploration for data augmentation for robust augmentation strategies such as Mixup. To tackle such challenges in the financial domain, we propose `SH-Mix`: Saliency-guided Hierarchical Mixup augmentation technique for multimodal financial prediction tasks. `SH-Mix` combines multi-level embedding mixup strategies based on the contribution of each modality and context subsequences. Through extensive quantitative and qualitative experiments on financial earnings and conference call datasets consisting of text and speech, we show that `SH-Mix` outperforms state-of-the-art methods by $3-7\%$. Additionally, we show that `SH-Mix` is generalizable across different modalities and models.

Keywords: Multimedia Document Processing, Social Media Processing, Tools, Systems, Applications

1. Introduction

Financial risk modelling is of great interest to capital market participants for making sound investment decisions comprising tasks like price forecasting and movement/volatility prediction which are essential to designing profitable trading strategies. Nevertheless, forecasting trends in these valuations is a complex task due to the inherent kinetic characteristics and volatility of the stock market.

Recent works establish the effectiveness of incorporating multimodal data from diverse sources like financial news (Hu et al., 2018), social media (Tabari et al., 2018) and financial documents (Mathur et al., 2022a) over conventional statistical methods using historical price data (Zheng et al., 2019; Ariyo et al., 2014a; Wu et al., 2022). Financial conference calls are one such rich information source consisting of textual, auditory and visual cues (Price et al., 2012; Brockman et al., 2017), and exhibit correlation with the involved firms' stock prices (Irani, 2004; Bowen et al., 2000). Examples include earnings calls, mergers and acquisitions calls, and monetary policy briefings which typically feature spoken content from senior executives and offer valuable insights into a company's performance. Including vocal cues along with explicit textual information enriches the learned representations with information about tonality, intonations, and pitch, which serve as indicators of underlying emotions and sentiment of the speaker, thereby

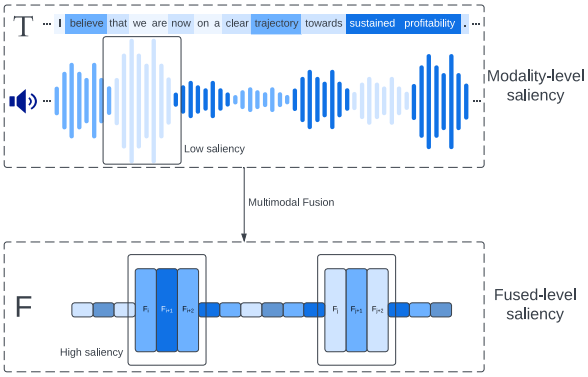


Figure 1: Example from a financial earnings call showing that not every part of the input has same relevance for price movement forecasting task. More salient spans are colored darker. There are evident variations in the degree of saliency observed within distinct hierarchical levels, specifically at the modality-level and the fused-level.

*Equal Contribution

LREC-COLING 2024, pages 14285–14297
20-25 May, 2024.

also making them contextually enhanced. Recent studies use multimodal fusion-based mechanisms to harness the contextual information for individual modalities within these calls, while simultaneously incorporating inter-modal relationships (Qin and Yang, 2019; Sawhney et al., 2021; Li et al., 2020; Mathur et al., 2022b). However, there is a scarcity of these real-world multimodal financial datasets, like financial earnings calls which can be as low as four calls per year (Chen et al., 2018). Additional challenges arise in the form of source variations and the need for meticulous annotation. To mitigate this scarcity, we explore Mixup (Zhang et al., 2018) as a data augmentation technique due to its established effectiveness in improving generalization ability in limited data domains (both unimodal and multimodal) (Chidambaram et al., 2022; Zhao et al., 2023; Lin and Hu, 2023; Liu et al., 2023).

Financial data, including conference and policy calls, comprises extensive long-form content, featuring extended audio-visual recordings (Behre et al., 2022), lengthy textual transcripts, typically containing over 5,000 words of text (Koval et al., 2023). Such large data streams contain certain salient segments that have the majority of informational essence (Wilmot and Keller, 2021). Saliency-aware Mixup techniques exploit this saliency information to mix only the most relevant parts of the input reducing noise by considering the discriminative features of the input while also preserving its local structure (Lee et al., 2022; Kim et al., 2020a; Sawhney et al., 2022a). It captures span-level saliencies in unimodal inputs and mixes them by transporting the salient span of one image/audio to another. These approaches utilize saliency from a unimodal perspective and do not consider cross-modal dependencies. As shown in Figure 1, multimodal data streams may contain certain important (salient) aspects to the model's predictions - both locally at the modality level and globally at the fused level, creating a hierarchical structure which is not fully captured by existing Mixup techniques.

Building on these gaps, we propose SH-Mix: Saliency-Aware Hierarchical Multimodal Mixup, a novel hierarchical architecture building on Mixup incorporating saliency information from the underlying data leveraging gradient-based measures. At the local-level, individual modality mixup is conducted based on modality-specific saliency. Subsequently, a second global-level is introduced, wherein these modality-specific representations are fused through an attention-based weighting mechanism, to obtain an abstract multimodal representation. At the global-level, saliency-based mixup is employed on these fused embeddings enabling it to capture the cross-modal correlations and inter-dependencies. Incorporated with a neural multimodal-fusion base, SH-Mix shows significant performance improvement compared to the state-of-the-art approaches. Our contributions are as follows:

- We introduce SH-Mix[1], a novel data augmentation strategy for multimodal financial data leveraging modality-level and fused-level saliency (**§3**).
- Through extensive quantitative (**§5.1**) and exploratory (**§5.2**, **§5.4**) experiments on real-world tasks with insufficient training data, such as financial prediction using conference calls, we show that SH-Mix outperforms the existing state of the art by 3-7%.
- Finally, we demonstrate the general applicability of our approach by presenting SH-Mix as a general Mixup framework for multimodal sequence learning through supplementary experiments on tasks from other domains like sarcasm detection and sentiment analysis, and different base models, on data comprising audio, text and visual sequences.(**§5.5**).

2. Related Work

AI in Finance Traditional approaches to financial forecasting employ numerical price-based data to predict future price volatility/movement. These include areas like stock market (Ariyo et al., 2014b; Rundo et al., 2019), cryptocurrencies (K et al., 2022; Iqbal et al., 2021) and currency exchange market (Kamruzzaman and Sarker, 2003). These approaches usually employ time-series models like ARIMA (Ariyo et al., 2014b) and GARCH (Bollerslev, 1986). Recently, textual, acoustic and visual signals fetched from social media platforms and web searches have been used for forecasting (Xu and Cohen, 2018; Sawhney et al., 2022b). Such signals are able to capture the underlying investor sentiment associated with financial security beyond just numerical data thus improving the forecasting ability of the model.

Multimodal Learning in Finance Recent multimodal learning advances have granted investors access to substantial structured and unstructured multimodal financial data for forecasting (Jiang, 2021). Anecdotal evidence highlights the relevance of non-verbal cues like vocal tone, emotional indicators and language complexity in relation to financial trading (Cao, 2022; Li et al., 2016b; Jiang and Pell, 2017). Studies by Qin and Yang (2019); Sawhney et al. (2020) have exploited multimodal data to predict both price volatility and movement. Mathur et al. (2022b) employed audio, visual, and textual

[1]Our code is released at https://github.com/gtfintechlab/shmix

cues from MPC calls to forecast price changes and volatility. The impact of augmentation methods, such as Mixup, during multimodal financial data training remains underexplored.

Mixup (Zhang et al., 2018) is a popular augmentation technique that interpolates two examples along with their corresponding labels. Existing work shows that Mixup performs well on tasks spanning vision, speech, and text (Meng et al., 2021; Chang et al., 2021; Verma et al., 2019; Chhabra et al., 2023; Sawhney and Neerkaje, 2022). Mixup strategies which operate on sequential data such as speech and/or text fail to preserve the locality of inputs while mixing at the input space. Recent work on saliency-based Mixup (Kim et al., 2020b; Ma et al., 2022) look to address this problem by mixing the most important contiguous spans over the raw inputs. Such saliency-based approaches when applied to sequential data, have only been explored in the context of unimodality, such as text (Kong et al., 2022; Yoon et al., 2021) and speech (Sawhney et al., 2022a). Mixup has also lately shown promise in multimodal setups (So et al., 2022; Hao et al., 2023; Meng et al., 2021; Zhou et al., 2023). Zhao et al. (2023) introduce a method which generates new virtual modalities from the mixed token-level representation of raw modalities. However, there is a gap in leveraging the salient components at the modality-level and sequence-level during Mixup, which is addressed by SH-Mix (**§3.4**).

3. Methodology

3.1. Background

Mixup generates virtual samples for training by a convex interpolation of training samples. Given two training samples x_i and x_j and their corresponding labels y_i and y_j, we generate synthetic sample $\tilde{x}$ and the corresponding mixed label $\tilde{y}$ as

$$\tilde{x} = mix(x_i, x_j) = \lambda x_i + (1-\lambda)x_j \quad (1)$$

$$\tilde{y} = mix(y_i, y_j) = \lambda y_i + (1-\lambda)y_j \quad (2)$$

$\lambda \in [0,1]$ is the mixing ratio (Zhang et al., 2018). In particular, for discrete classification settings, y_i and y_j are one-hot encoded labels.

Saliency measures the contribution of input / hidden representation features in predicting a specific output class. Gradient-based methods for saliency computation (Simonyan et al., 2014; Li et al., 2016a) are used during training to find features contributing the most towards the prediction. We compute the saliency for an input / hidden representation $Z = [z_1, z_2, \ldots, z_n]$ by computing its gradient with respect to the classification loss $\mathcal{L}$.

$$\text{sal}(Z;\mathcal{L}) = \frac{\partial \mathcal{L}}{\partial Z} = \left[\frac{\partial \mathcal{L}}{\partial z_1}, \frac{\partial \mathcal{L}}{\partial z_2}, \ldots, \frac{\partial \mathcal{L}}{\partial z_n}\right] \quad (3)$$

3.2. Problem Formulation

Given an example $\mathcal{X} = [\mathcal{R}^1, \mathcal{R}^2, \ldots, \mathcal{R}^N]$ where $\mathcal{X}$ is a composite multimodal sequence comprising of N distinct modalities. Each modality $\mathcal{R}^i$ comprises of a temporal sequence of its raw inputs $\mathcal{R}^i = [r^i_1, r^i_2, \ldots, r^i_n]$ where n is the length of each modality's sequence. All the modalities are temporally aligned.

Following Xu and Cohen (2018), we define *price movement prediction* as a binary classification task which uses the multimodal input $\mathcal{X}$ to predict the price movement for the associated firm's stock's closing price over a period of τ days following the conference call. We define the movement label $y_{d-\tau,d}$ as

$$y_{d-\tau,d} = \begin{cases} 1 & p_d > p_{d-\tau} \\ 0 & p_d < p_{d-\tau} \end{cases} \quad (4)$$

where p_d is the closing price on the day d.

3.3. ADMF: Attention-Driven Multimodal Fusion Architecture

Utilizing low-level modality-specific specialized transformers, such as BERT (Devlin et al., 2019), ViT (Dosovitskiy et al., 2021), and AST (Gong et al., 2021) on raw inputs enables independent processing of each modality at the utterance level, facilitating the capture of modality-specific patterns and yielding contextually rich representations. In line with existing works by Tsai et al. (2019); Mathur et al. (2022b), we first convert the raw modality data into embeddings via low-level transformers (ϕ_i) corresponding to each modality to get $\mathcal{M}^i = [m^i_1, m^i_2, \ldots, m^i_n]$ where $m^i_j = \phi_i(r^i_j)$. We then use an attention-based fusion mechanism that captures the dependency between the modalities. The attention weights W'_i for a modality are computed via softmax normalization as

$$W_i = \frac{e^{\mathcal{M}^i A^i + b^i}}{\sum_{j=1}^{k} e^{\mathcal{M}^j A^j + b^j}} \quad \forall i \in [1,2,\ldots,N] \quad (5)$$

$$W'_i = \frac{W_i}{\sum_{k=1}^{N} W_k} \quad \forall i \in [1,2,\ldots,N] \quad (6)$$

where A^i and b^i represent the attention layer weights learned during the training.

The attention weights are used to weigh the features corresponding to each modality by multiplying these with the respective embeddings to obtain the attended inputs. This adaptive attention-based weighing mechanism enables the model to selectively focus on the most informative modalities along with temporal dependencies (Hori et al.,

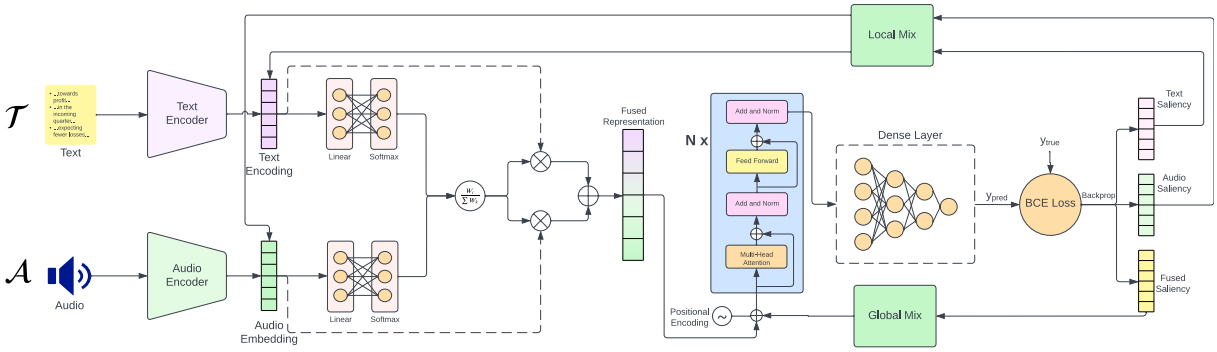


Figure 2: **SH-Mix Overview:** Input text $\mathcal{T}$ and audio $\mathcal{A}$ are encoded to yield respective embeddings. Attention weights for each modality are extracted, followed by fusion through weighted summation. The fused representation is fed to transformer block and a dense layer, to obtain the loss. Saliencies are computed for text, audio, and the fused representation via backpropagation, as detailed in section 3.1. These saliencies drive Local-Mix and Global-Mix, resulting in two sets of mixed inputs.

2017; Yan et al., 2020). These are further combined by additive fusion to obtain the intermediate fused multimodal embedding. This is augmented with positional embedding (POS) by addition to obtain the final fused embedding F as follows

$$F = \sum_{j=1}^{N} \mathcal{M}^j W'_j + \text{POS} \tag{7}$$

We use a transformer encoder which employs multi-headed self-attention (Vaswani et al., 2017) along with a feed-forward network to obtain the encoded representations for the input fused embedding F. Average pooling is applied to the output of the transformer before passing through two dense layers (MLP) to produce the output $y = \text{MLP}(F) = f_\theta(\mathcal{X})$, where $f_\theta(\cdot)$ represents the complete model architecture with parameters θ.

3.4. SH-Mix: Components

Given multimodal examples $\mathcal{X}_A$ and $\mathcal{X}_B$, we embed each modality as: $\mathcal{X}_A = [\mathcal{M}_A^1, \mathcal{M}_A^2, \ldots, \mathcal{M}_A^N]$ and $\mathcal{X}_B = [\mathcal{M}_B^1, \mathcal{M}_B^2, \ldots, \mathcal{M}_B^N]$. To compute saliency information at the global and local levels, we also perform a forward pass on the unmixed inputs to obtain an initial unmixed loss $\mathcal{L}_{org}$ (**§3.5**).

Local-Mix To capture the most important aspects of a given modality, and keep the mixed sample more closely related to the output, modality-specific Mixup is applied to generate the mixed multimodal input $\tilde{\mathcal{X}}$. We find the saliency S_A^i of modality $\mathcal{M}_A^i$ and S_B^i of $\mathcal{M}_B^i$ as,

$$S_A^i = \text{sal}(\mathcal{M}_A^i; \mathcal{L}_{org}) \; ; S_B^i = \text{sal}(\mathcal{M}_B^i; \mathcal{L}_{org})$$

where $S_A^i = [(s_1^i)_A, (s_2^i)_A, \ldots, (s_n^i)_A]$ and $S_B^i = [(s_1^i)_B, (s_2^i)_B, \ldots, (s_n^i)_B]$. For $(m_j^i)_A \in \mathcal{M}_A^i$ and $(m_j^i)_B \in \mathcal{M}_B^i$, we find positions of the k_i greatest values in $(s_j^i)_A$ and the k_i least values in $(s_j^i)_B$.

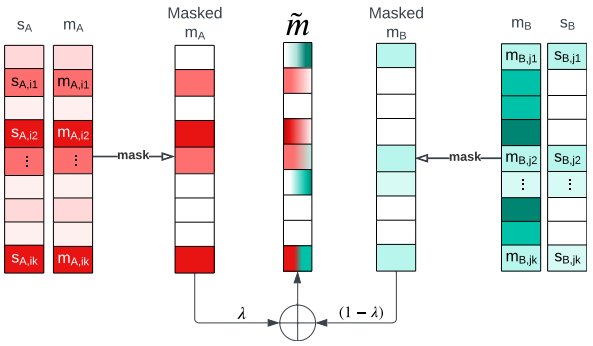


Figure 3: **Local-Mix**: Given any audio/text utterance m_A and m_B for input samples A and B, we mix the most salient portions of m_A with the least salient portions of m_B while zeroing out the remaining features to obtain the mixed utterance $\tilde{m}$. More salient features are darker.

Here $k_i = \delta_{\text{loc}} \cdot p_i$, where p_i is the number of features in the embedding corresponding to the i^{th} modality and δ_{loc} is a hyperparameter for controlling the local Mixup threshold. We perform Mixup on the features present at these positions. We define binary masks $(mask_j^i)_A$ and $(mask_j^i)_B$ of size p_i. $(mask_j^i)_A$ is 1 at k_i most salient positions of $(m_j^i)_A$ and $(mask_j^i)_B$ is 1 at k_i least salient positions of $(m_j^i)_B$. These masks are used to zero out the features that are not involved in Mixup. We then define Local-Mix as,

$$\begin{aligned} \tilde{m}_j^i &= \lambda_{loc}(m_j^i)_A \odot (mask_j^i)_A \\ &\quad + (1-\lambda_{loc})(m_j^i)_B \odot (mask_j^i)_B \end{aligned} \tag{8}$$

$$\tilde{\mathcal{M}}^i = [\tilde{m}_1^i, \tilde{m}_2^i, \ldots, \tilde{m}_n^i] \tag{9}$$

$$\begin{aligned} \tilde{\mathcal{X}} &= \text{locmix}(\mathcal{X}_A, \mathcal{X}_B, \mathcal{L}_{org}) \\ &= [\tilde{\mathcal{M}}^1, \tilde{\mathcal{M}}^2, \ldots, \tilde{\mathcal{M}}^N] \end{aligned} \tag{10}$$

where λ_{loc} is the local mixing ratio and is sampled from a beta distribution. The labels associated with examples $\mathcal{X}_A$ and $\mathcal{X}_B$ i.e. y_A and y_B are also

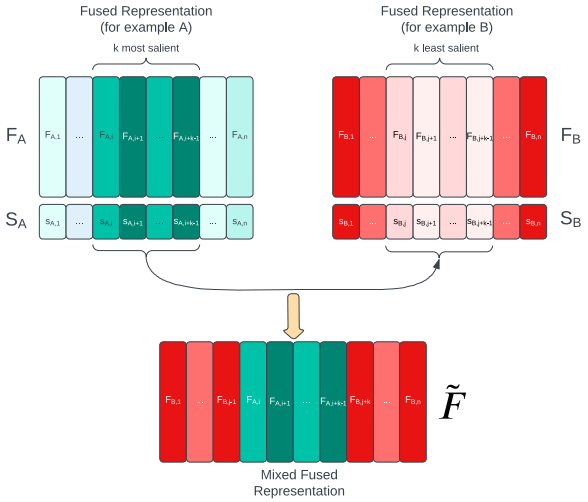


Figure 4: **Global-Mix**: Given fused representations F_A and F_B of input samples A and B, we replace the most salient span in F_A with the least salient span in F_B to obtain the mixed fused representation $\tilde{F}$. More salient utterances are colored darker.

mixed to obtain the mixed label $\tilde{y}_{\text{loc}}$.

$$\tilde{y}_{\text{loc}} = \lambda_{\text{loc}} y_A + (1 - \lambda_{\text{loc}}) y_B \tag{11}$$

Global-Mix Span Mixup approaches have shown to be more effective in preserving the locality and maintaining the structural coherence of the inputs (Yun et al., 2019); synergistically integrated with saliency-driven methods it has shown to effectively utilize the ingrained local statistics in the data (Kim et al., 2020a). Global-Mix uses the fused representation of the data to enable efficient cross-modal information exchange and leverage these contextual embeddings. For each example $\mathcal{X}_A$ and $\mathcal{X}_B$, we compute utterance-level saliency for their fused representation $F_A = [u_1, u_2, \ldots, u_n]$ and $F_B = [u'_1, u'_2, \ldots, u'_n]$ obtained using ADMF, given as,

$$S_A = \text{sal}(F_A; \mathcal{L}_{org}) \;\; ; \;\; S_B = \text{sal}(F_B; \mathcal{L}_{org}) \tag{12}$$

where $S_A = [(s_1)_A, (s_2)_A, \ldots, (s_n)_A]$ and $S_B = [(s_1)_B, (s_2)_B, \ldots, (s_n)_B]$. We compute the span-level saliency of span p to q as the sum of L_1 norm of the saliencies in the span, given as,

$$s_A[p;q] = \sum_{k=p}^{q} \|(s_k)_A\|_{L_1} \tag{13}$$

$$s_B[p;q] = \sum_{k=p}^{q} \|(s_k)_B\|_{L_1} \tag{14}$$

As shown in Figure 4, on a span of length $l = \lambda_{\text{glob}} \cdot n$ between inputs $\mathcal{X}_A$ and $\mathcal{X}_B$, we replace the most salient span $[i; i+l-1]$ in F_A with the least salient span $[j; j+l-1]$ in F_B as,

$$i = \arg\max_i \; s_A[i; i+l-1] \tag{15}$$

$$j = \arg\min_j \; s_B[j; j+l-1] \tag{16}$$

$$\tilde{F} = \text{globmix}(F_A, F_B, \mathcal{L}_{org}) = \begin{cases} u'_k & k \notin [j, j+1, \ldots j+l-1] \\ u_{i+k-j} & k \in [j, j+1, \ldots j+l-1] \end{cases} \tag{17}$$

The labels associated with examples $\mathcal{X}_A$ and $\mathcal{X}_B$ i.e. y_A and y_B are mixed with the global mixing ratio λ_{glob} to obtain the mixed label $\tilde{y}_{\text{glob}}$,

$$\tilde{y}_{\text{glob}} = \lambda_{\text{glob}} \cdot y_A + (1 - \lambda_{\text{glob}}) \cdot y_B \tag{18}$$

3.5. SH-Mix Traning Objective

The unmixed inputs $[\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_N]$ are passed through the model $f_\theta(\cdot)$ to get the corresponding logit y'_{org}. $\mathcal{L}_{\text{org}}$ is computed as the binary cross entropy (BCE) loss between the predicted logit y'_{org} and the ground truth y_{org}.

$$y'_{\text{org}} = f_\theta(X) \quad ; \quad \mathcal{L}_{\text{org}} = \text{BCE}(y'_{\text{org}}, y_{\text{org}}) \tag{19}$$

We perform backpropagation on the loss obtained above to get gradients w.r.t. all the input modalities $\mathcal{M}^1$ through $\mathcal{M}^N$ as well as the fused embedding F described in equation 7.

For any two unmixed input examples $\mathcal{X}_A$ and $\mathcal{X}_B$, we perform Local-Mix to get the mixed input $\tilde{\mathcal{X}} = \text{locmix}(\mathcal{X}_A, \mathcal{X}_B, \mathcal{L}_{org})$ and $\tilde{y}_{\text{loc}}$. We pass $\tilde{\mathcal{X}}$ through the model and compute the BCE loss with the output y'_{loc} to get the loss $\mathcal{L}_{\text{loc}}$

$$y'_{\text{loc}} = f_\theta(\tilde{\mathcal{X}}) \quad ; \quad \mathcal{L}_{\text{loc}} = \text{BCE}\left(y'_{\text{loc}}, \tilde{y}_{\text{loc}}\right) \tag{20}$$

We similarly perform Global-Mix on $\mathcal{X}_A$ and $\mathcal{X}_B$ to get mixed fused embedding $\tilde{F} = \text{globmix}(\mathcal{X}_A, \mathcal{X}_B, \mathcal{L}_{org})$ and mixed label $\tilde{y}_{\text{glob}}$. We now pass $\tilde{F}$ into the transformer encoder to get y'_{glob} and find the BCE loss with $\tilde{y}_{\text{glob}}$ to get the Global-Mix loss $\mathcal{L}_{\text{glob}} = \text{BCE}\left(y'_{\text{glob}}, \tilde{y}_{\text{glob}}\right)$. The three losses are combined to get a single weighted loss

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{org}} + \beta \cdot \mathcal{L}_{\text{loc}} + \gamma \cdot \mathcal{L}_{\text{glob}} \tag{21}$$

4. Experiments

4.1. Datasets

Multimodal Multi-Speaker Merger & Acquisition Call (M&A Calls) (Sawhney et al., 2021) consists of 812 M&A conference calls between 2016 to 2020. The data for each call comprises of the text transcript and the aligned audio recording of the

Model	M&A Calls Dataset						MAEC Dataset					
	$\tau = 3$		$\tau = 7$		$\tau = 15$		$\tau = 3$		$\tau = 7$		$\tau = 15$	
Metric	$F1_3$	MCC_3	$F1_7$	MCC_7	$F1_{15}$	MCC_{15}	$F1_3$	MCC_3	$F1_7$	MCC_7	$F1_{15}$	MCC_{15}
MLP	0.52	0.10	0.57	0.17	0.48	-0.04	0.50	0.09	0.55	0.13	0.55	0.10
LSTM	0.58	0.15	0.54	0.12	0.51	0.11	0.54	0.16	0.50	0.01	*0.56*	0.12
MulT	0.57	0.18	0.57	0.18	0.52	0.12	0.55	0.13	0.55	0.10	0.54	0.09
MDRM	0.57	*0.20*	0.58	0.19	0.46	0.11	**0.60**	**0.20**	0.54	0.11	*0.56*	*0.13*
M3ANet	*0.59*	0.18	0.58	0.17	0.50	0.13	0.56	0.13	0.54	0.09	0.55	0.10
M3ANet + Mixup	0.58	0.17	0.60	*0.21*	*0.57*	*0.15*	0.56	0.12	*0.56*	*0.12*	*0.56*	0.12
PISA	*0.59*	0.19	*0.61*	0.18	0.55	0.14	0.57	0.15	0.55	0.11	0.54	0.12
SH-Mix (Ours)	**0.66***	**0.32***	**0.63***	**0.30***	**0.63***	**0.26***	*0.58*	*0.18*	**0.58***	**0.17***	**0.56***	**0.13***

Table 1: Performance comparison of SH-Mix with both non-transformer and transformer-based baselines for M&A calls and MAEC dataset for price movement prediction τ days after the call where $\tau \in \{3, 7, 15\}$. $*$ shows statistically significant improvements $(p < 0.005)$ over PISA under Wilcoxon's Signed Rank Test. **Bold**, *italics* shows the **best**, *second best* performance.

call along with the speaker ID associated with each utterance. We encode text transcripts using BERT (Devlin et al., 2019) and process audio recording using OpenSMILE[2]. We use the train-test split as released with the dataset.

Multimodal Aligned Earnings Call (MAEC) (Li et al., 2020) contains aligned text transcripts and associated audio recordings from the earnings calls of S&P 1500 companies between 2015 to 2018. For each example, the text is procesed using BERT encoder and the corresponding audio recording is processed using *Praat* (Boersma and Van Heuven, 2001). We use 860 data samples with a training:validation:test split ratio as $60 : 10 : 30$, same as the released dataset.

4.2. Baseline Models

We compare SH-Mix with several conventional and contemporary multimodal and mixup-based baselines:

MLP The encoding corresponding to each modality is averaged along the time axis and simply concatenated before passing through a vanilla multi-layer perceptron network.

LSTM Inputs multimodal time series to individual LSTMs (Hochreiter and Schmidhuber, 1997) and averages before making the final prediction using a dense layer.

MulT (Tsai et al., 2019) fuses multimodal sequences using directional pairwise cross-modal transformers followed by sequence models for predictions.

MDRM (Qin and Yang, 2019) uses a contextual BiLSTM (Poria et al., 2017) to derive context-aware unimodal sequence representations, which are then fused together using another layer of BiLSTM to extract multimodal inter-dependencies.

M3ANet (Sawhney et al., 2021) employs attention weights for multimodal fusion, capturing inter-dependency between modalities utilizing multi-head attention to model long-range dependencies and incorporate local and global contextual information.

M3ANet + Mixup Simple linear Mixup (Zhang et al., 2018) is performed at the individual modality level before feeding it to M3ANet model. The mixing ratio is sampled out of a beta distribution.

PISA (Sawhney et al., 2022a) Current state-of-the-art among saliency-based approaches. Applies a portion-wise Mixup method that leverages the hyperbolic space to model complex hierarchies in the data. Since PISA operates on a single modality, we fuse the individual modalities before feeding the inputs to the model.

4.3. Training Setup

The architecture of ADMF consists of a hidden layer of size 32 with ReLU activation. We use a transformer block with a feed-forward layer size of 64 with 3 attention heads for M&A calls and 4 attention heads for the MAEC dataset. The batch size for all our experiments is 64. We have used Keras[3] for our implementation. We report the weighted F-1 score and MCC as the mean of 10 independent runs.

We tune the hyper-parameters using Optuna[4] framework. The Mixup-related parameters are tuned by sampling from the following ranges: Global-Mix ratio $\lambda_{\text{glob}} \in [0.1, 0.9]$, Local-Mix threshold $\delta_{\text{loc}} \in [0.1, 0.9]$, $\alpha \in [0, 1]$, $\beta \in [0, 1]$, $\gamma \in [0, 1]$, and the initial learning rate $\in [1e-4, 2e-3]$ optimized with Adam optimizer. We use TPE (Tree-structured Parzen Estimator) algorithm (Bergstra et al., 2011) as the sampling strategy for hyper-

[2] https://github.com/audeering/opensmile

[3] https://keras.io/

[4] https://optuna.org/

Hyperparameter	M&A Dataset			MAEC Dataset		
	$\tau = 3$	$\tau = 7$	$\tau = 15$	$\tau = 3$	$\tau = 7$	$\tau = 15$
Global-Mix ratio (λ_{glob})	0.38	0.26	0.56	0.23	0.32	0.32
Local-Mix threshold (δ_{loc})	0.80	0.50	0.53	0.59	0.65	0.55
Loss coefficient for $\mathcal{L}_{\text{org}}(\alpha)$	0.20	0.87	0.10	0.13	0.58	0.15
Loss coefficient for $\mathcal{L}_{\text{loc}}(\beta)$	0.16	0.19	0.22	0.10	0.62	0.29
Loss coefficient for $\mathcal{L}_{\text{glob}}(\gamma)$	0.14	0.12	0.14	0.66	0.26	0.17
Initial learning rate	$7e-4$	$1e-4$	$1.6e-3$	$1.1e-3$	$9e-4$	$1e-3$

Table 2: Hyperparameter optimization results

parameter tuning. Table 2 provides the details for the best hyperparameters obtained for our model for price movement prediction task on M&A calls and MAEC dataset τ days after the call where $\tau \in \{3, 7, 15\}$.

4.4. Infrastructure and Compute details

Our model has 105,932 trainable parameters for M&A calls dataset and 48,710 trainable parameters for MAEC dataset. To generate text embeddings, we use a pre-trained BERT base model (uncased), which has 110 million parameters. We run our experiments on one NVIDIA A2 GPU which has 16 GB memory. The total number of GPU hours are 816 hours across all the experiments.

4.5. Evaluation Metrics

Following prior work (Mathur et al., 2022b; Sawhney et al., 2021) and given the data imbalance, we use weighted F1-score and Matthews correlation coefficient (MCC) (Matthews, 1975). MCC also mitigates the impact of class distribution skew and is invariant to the class labels.

5. Results and Discussion

5.1. Performance Comparison

Table 1 shows our model's performance on the financial datasets against the baselines. Models like MLP and LSTM perform feature aggregation over time series/modalities, and MDRM (Qin and Yang, 2019) uses hierarchical multimodal fusion on top of a contextual LSTM (Poria et al., 2017), giving a better representation of individual modalities with temporal dependencies. MulT (Tsai et al., 2019) attends to low-level features and captures long-range dependencies across modalities, thus exhibiting improved performance over these feature aggregation based models. Attention-based transformer models like M3ANet (Sawhney et al., 2021) are able to focus on the relevant parts of an input sequence while also preserving long-range dependencies, thus outperforming the other baselines (Wu et al., 2021; Tsai et al., 2019; Xu et al., 2023; Mathur et al., 2022b). Addition of Mixup to M3ANet further boosts its performance corroborating Mixup's regularization and data augmentation benefits in improving generalization in cases of feature diversity (Carratino et al., 2022). The incorporation of saliency information, utilized in the hyperbolic space for capturing complex geometries, to selectively perform portion-wise mixup in PISA yields performance improvement. Our proposed SH-Mix shows significant performance improvement over existing multimodal baselines, surpassing state-of-the-art by $3-7\%$. These observations verify our hypothesis that SH-Mix is able to select relevant multimodal features across different modalities and timestamps due to hierarchical saliency-guided components that interpolate discriminative temporal spans closely related to the prediction. Further, it also preserves local features key to modeling sequential information similar to Yoon et al. (2021); Kim et al. (2020a). We attribute the gains to the saliency-driven modality-specific nature of interactions in SH-Mix arising due to fine-grained local and global Mixup at the embedding level.

5.2. Ablation: Impact of Saliency

In Table 3, we conduct an ablation analysis to investigate the contribution of local and global saliency Mixup. We observe that omitting local saliency lowers performance as the model is unable to utilize fine-grained, modality-specific features. Removing global saliency Mixup also leads to performance degradation due to the loss of context-aware fused representations which are crucial for capturing high-level cross-modal dependencies. SH-Mix, combining both global and local saliency, achieves superior performance as it effectively exploits the locally discriminative features in each modality while also capturing broader contextual relationships across the input token sequence. This combined approach brings diversity and qualitative enrichment, resulting in improved generalization.

Saliency Component	M&A Calls		MAEC	
	$F1_3$	MCC_3	$F1_3$	MCC_3
All (SH-Mix)	**0.66***	**0.32***	**0.58***	**0.18***
(×) Saliency (Local)	0.58	0.18	0.55	0.11
(×) Saliency (Global)	0.59	0.19	0.57	0.14

Table 3: Ablation study covering performance impact of the removal of each kind of saliency component i.e. local and global level Mixup ($\tau = 3$ days). The combined model with both components shows the best results. $*$ shows statistically significant improvements ($p < 0.005$) over other configurations under Wilcoxon's Signed Rank Test. **Bold** shows the **best** performance.

Modality	M&A Calls		MAEC	
	$F1_3$	MCC_3	$F1_3$	MCC_3
Only Audio (A)	0.52	0.07	0.53	0.08
Only Text (T)	0.59	0.19	0.54	0.10
Audio + Text (AT)	**0.66***	**0.32***	**0.58***	**0.18***

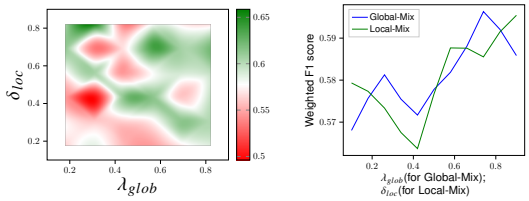
Table 4: Impact of Modality: Text-based (T) unimodal model gives better performance compared to audio-based (A) unimodal model due to noise in acoustic inputs. Combining both modalities (T+A) gives the best performance.$*$ shows statistically significant improvements ($p < 0.005$) over other configurations under Wilcoxon's Signed Rank Test. **Bold** shows the **best** performance.

5.3. Impact of Modality

Table 4 compares the multimodal saliency-guided Mixup model (AT) with its unimodal counterparts (A, T). We observe that SH-Mix (T) outperforms SH-Mix (A), which can be attributed to the inherent noise and variability in audio data (Mathur et al., 2022a; Chorowski et al., 2015), showing that saliency information extracted from an explicit semantic representation of text is more relevant to the task. We note that SH-Mix significantly outperforms its unimodal variants as multimodality helps to leverage complementary strengths of text utterances and acoustic features, similar to prior works Tsai et al. (2019); Mathur et al. (2022b).

5.4. Impact of Augmentation Strength

We observe optimal performance surfaces in the region characterized by both high δ_{loc} and high λ_{glob} (Figure 5a). A higher threshold δ_{loc} effectively selects the most salient, discriminative features for Local-Mix and a higher mixing ratio λ_{glob} prioritizes longer informative spans (Sawhney et al., 2022a), enhancing global Mixup quality by incorporating rich cross-modal contextual information. Consequently, reduced values for both exhibit a discernible performance drop. Figure 5b reaffirms



(a) Joint Impact (b) Individual Impact

Figure 5: Joint and individual impact of Global-Mix mixing ratio (λ_{glob}) and Local-Mix threshold (δ_{loc}) on SH-Mix's performance on M&A calls. Higher λ_{glob} and δ_{loc} boost performance for both cases.

this trend, wherein increasing both λ_{glob} and δ_{loc} independently positively impact model performance. Note that other regions with optimal performance also exist as is evident from Figure 5a, e.g. high δ_{loc} and low λ_{glob}, which is also the configuration captured during hyperparameter tuning.

Model Setup	MUStARD		CMU-MOSI	
	F1	MCC	F1	MCC
ADMF	0.65	0.28	0.75	0.50
ADMF + Mixup	0.68	0.35	0.76	0.52
SH-Mix (Ours)	**0.71***	**0.41***	**0.78***	**0.54***

Table 5: Modality agnostic generalizability: Performance of SH-Mix on audio, text, and video modalities in MUStARD and CMU-MOSI. SH-Mix outperforms ADMF and ADMF + Mixup.$*$ shows statistically significant improvements ($p < 0.005$) over other configurations under Wilcoxon's Signed Rank Test. **Bold** shows the **best** performance.

5.5. General Applicability of SH-Mix

To gauge the adaptability of SH-Mix to diverse applications, we apply it to sarcasm detection and sentiment analysis binary classification tasks on the MUStARD (Castro et al., 2019) and CMU-MOSI (Zadeh et al., 2016) dataset respectively. Both of these comprise audio, visual and textual sequences. Table 5 shows the effectiveness of SH-Mix over general fusion-based models and vanilla Mixup techniques. It also displays a model-agnostic behaviour, exhibiting performance improvement when applied over different neural baselines (see Table 6). By synergizing the hierarchy with saliency, SH-Mix leverages discriminative modality-specific features and informative global spans, thus establishing it as a comprehensive hierarchical multimodal sequence learning Mixup framework.

Model Setup	M&A Calls		MAEC	
	F1	MCC	F1	MCC
MDRM	0.57	0.20	0.60	0.20
SH-Mix (MDRM)	**0.61***	**0.23***	**0.61***	**0.21***
M3ANet	0.58	0.17	0.56	0.12
SH-Mix (M3ANet)	**0.63***	**0.26***	**0.56***	**0.13***

Table 6: SH-Mix is generalizable across neural architectures like MDRM and M3ANet.$*$ shows statistically significant improvements ($p < 0.005$) over other configurations under Wilcoxon's Signed Rank Test. **Bold** shows the **best** performance.

6. Conclusion

Building on the current limitations in multimodal augmentation strategies, we introduced SH-Mix: a saliency-guided hierarchical Mixup technique for multimodal financial prediction tasks. SH-Mix combines multi-level embedding Mixup strategies based on the contribution of each modality and contextual subsequences. Experimental results show that it outperforms the existing state of the art by $3 - 7\%$. We further analyze the contribution of local and global levels of saliency, evaluate the impact of each modality and assess the impact of augmentation strength within SH-Mix. We also show that SH-Mix is generalizable across different modalities and neural models

7. Ethical Considerations and Limitations

Our research specifically hones in on conference calls in which companies make both transcripts and audio recordings publicly accessible.The data for M&A calls and Earnings conference calls is openly available for anyone to download. Our usage and storage of all the company data strictly adheres to privacy laws, with no collection of personal data or violations.

Limitations We acknowledge the presence of gender bias in our study caused by the speaker-level gender imbalance in the M&A calls and Earnings calls. We also acknowledge the presence of demographic bias in our study since the calls belong to companies in the United States of America and cannot be generalized to other geographies and non-native speakers. Also, our study is limited to English language motivating similar study on other multilingual calls.

Potential Risks It is also crucial to note that our work of exploratory research should not be treated as explicit financial advice. All investment-related decisions involve exposure to market risks and should only be taken following thorough evaluation.

8. Bibliographical References

Adebiyi A. Ariyo, Adewumi O. Adewumi, and Charles K. Ayo. 2014a. Stock price prediction using the arima model. In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, pages 106–112.

Adebiyi A. Ariyo, Adewumi O. Adewumi, and Charles K. Ayo. 2014b. Stock price prediction using the arima model. In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, pages 106–112.

Piyush Behre, S.S. Tan, Padma Varadharajan, and Shuangyu Chang. 2022. Streaming punctuation for long-form dictation with transformers. *ArXiv*, abs/2210.05756.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 2546–2554, Red Hook, NY, USA. Curran Associates Inc.

Paul Boersma and Vincent Van Heuven. 2001. Speak and unspeak with praat. *Glot Int*, 5:341–347.

Tim Bollerslev. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Robert M. Bowen, Angela K. Davis, and Dawn A. Matsumoto. 2000. Do conference calls affect analysts' forecasts? *SSRN Electronic Journal*.

Paul Brockman, Xu Li, and S. McKay Price. 2017. Conference call tone and stock returns: Evidence from the stock exchange of hong kong. *Asia-Pacific Journal of Financial Studies*, 46(5):667–685.

Longbing Cao. 2022. Ai in finance: Challenges, techniques, and opportunities. *ACM Comput. Surv.*, 55(3).

Luigi Carratino, Moustapha Cissé, Rodolphe Jenatton, and Jean-Philippe Vert. 2022. On mixup regularization. *J. Mach. Learn. Res.*, 23(1).

Oscar Chang, Dung N Tran, and Kazuhito Koishida. 2021. Single-channel speech enhancement using learnable loss mixup. In *Interspeech*, pages 2696–2700.

Jason V Chen, Venky Nagar, and Jordan Schoenfeld. 2018. Manager-analyst conversations in earnings conference calls. *Review of Accounting Studies*, 23:1315–1354.

Parth Chhabra, Atula Tejaswi Neerkaje, Shivam Agarwal, Ramit Sawhney, Megh Thakkar, Preslav Nakov, and Sudheer Chava. 2023. Learning through interpolative augmentation of dynamic curvature spaces. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2108–2112.

Muthu Chidambaram, Xiang Wang, Yuzheng Hu, Chenwei Wu, and Rong Ge. 2022. Towards understanding the data dependency of mixup-style training. In *International Conference on Learning Representations*.

Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 577–585, Cambridge, MA, USA. MIT Press.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio spectrogram transformer. In *Interspeech 2021*. ISCA.

Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. 2023. Mixgen: A new multi-modal data augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 379–389.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

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 *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4203–4212.

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, page 261–269, New York, NY, USA. Association for Computing Machinery.

Mahir Iqbal, Muhammad Shuaib Iqbal, Fawwad Hassan Jaskani, Khurum Iqbal, and Ali Hassan. 2021. Time-series prediction of cryptocurrency market using machine learning techniques. *EAI Endorsed Transactions on Creative Technologies*, 8(28).

Afshad J. Irani. 2004. The effect of regulation fair disclosure on the relevance of conference calls to financial analysts. *Review of Quantitative Finance and Accounting*, 22(1):15–28.

Weiwei Jiang. 2021. Applications of deep learning in stock market prediction: Recent progress. *Expert Syst. Appl.*, 184(C).

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

Dhinakaran K, Baby Shamini P, Divya J, Indhumathi C, and Asha R. 2022. Cryptocurrency exchange rate prediction using arima model on real time data. In *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, pages 914–917.

J. Kamruzzaman and R.A. Sarker. 2003. Forecasting of currency exchange rates using ann: a case study. In *International Conference on Neural Networks and Signal Processing, 2003. Proceedings of the 2003*, volume 1, pages 793–797 Vol.1.

Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. 2020a. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

JangHyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. 2020b. Co-mixup: Saliency guided joint mixup with supermodular diversity. In *International Conference on Learning Representations*.

Fanshuang Kong, Richong Zhang, Xiaohui Guo, Samuel Mensah, and Yongyi Mao. 2022. Dropmix: A textual data augmentation combining

dropout with mixup. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 890–899.

Ross Koval, Nicholas Andrews, and Xifeng Yan. 2023. Forecasting earnings surprises from conference call transcripts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8197–8209, Toronto, Canada. Association for Computational Linguistics.

Sanghyeok Lee, Minkyu Jeon, Injae Kim, Yunyang Xiong, and Hyunwoo J. Kim. 2022. Sagemix: Saliency-guided mixup for point clouds. In *Advances in Neural Information Processing Systems*.

Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. CIKM '20, page 3063–3070, New York, NY, USA. Association for Computing Machinery.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016a. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Yelin Li, Junjie Wu, and Hui Bu. 2016b. When quantitative trading meets machine learning: A pilot survey. In *2016 13th International Conference on Service Systems and Service Management (ICSSSM)*, pages 1–6.

Ronghao Lin and Haifeng Hu. 2023. Adapt and explore: Multimodal mixup for representation learning.

Zichang Liu, Zhiqiang Tang, Xingjian Shi, Aston Zhang, Mu Li, Anshumali Shrivastava, and Andrew Gordon Wilson. 2023. Learning multimodal data augmentation in feature space. In *The Eleventh International Conference on Learning Representations*.

Avery Ma, Nikita Dvornik, Ran Zhang, Leila Pishdad, Konstantinos G. Derpanis, and Afsaneh Fazly. 2022. SAGE: saliency-guided mixup with optimal rearrangements. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 484. BMVA Press.

Puneet Mathur, Mihir Goyal, Ramit Sawhney, Ritik Mathur, Jochen Leidner, Franck Dernoncourt, and Dinesh Manocha. 2022a. DocFin: Multimodal financial prediction and bias mitigation using semi-structured documents. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1933–1940, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Puneet Mathur, Atula Neerkaje, Malika Chhibber, Ramit Sawhney, Fuming Guo, Franck Dernoncourt, Sanghamitra Dutta, and Dinesh Manocha. 2022b. Monopoly: Financial prediction from monetary policy conference videos using multimodal cues. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 2276–2285, New York, NY, USA. Association for Computing Machinery.

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.

Linghui Meng, Jin Xu, Xu Tan, Jindong Wang, Tao Qin, and Bo Xu. 2021. Mixspeech: Data augmentation for low-resource automatic speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7008–7012. IEEE.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

S. McKay Price, James S. Doran, David R. Peterson, and Barbara A. Bliss. 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4):992–1011.

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.

Francesco Rundo, Francesca Trenta, Agatino Luigi di Stallo, and Sebastiano Battiato. 2019. Machine learning for quantitative finance applications: A survey. *Applied Sciences*, 9(24).

Ramit Sawhney, Mihir Goyal, Prakhar Goel, Puneet Mathur, and Rajiv Ratn Shah. 2021. Multimodal multi-speaker merger & acquisition financial modeling: A new task, dataset, and neural

baselines. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6751–6762, Online. Association for Computational Linguistics.

Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. 2020. Multimodal multi-task financial risk forecasting. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 456–465, New York, NY, USA. Association for Computing Machinery.

Ramit Sawhney and Atula Tejaswi Neerkaje. 2022. Intermix: An interference-based data augmentation and regularization technique for automatic deep sound classification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3443–3447. IEEE.

Ramit Sawhney, Megh Thakkar, Vishwa Shah, Puneet Mathur, Vasu Sharma, and Dinesh Manocha. 2022a. PISA: PoIncaré Saliency-Aware Interpolative Augmentation. In *Proc. Interspeech 2022*, pages 2663–2667.

Ramit Sawhney, Megh Thakkar, Ritesh Soun, Atula Neerkaje, Vasu Sharma, Dipanwita Guhathakurta, and Sudheer Chava. 2022b. Tweet based reach aware temporal attention network for NFT valuation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6321–6332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*.

Junhyuk So, Changdae Oh, Yongtaek Lim, Hoyoon Byun, Minchul Shin, and Kyungwoo Song. 2022. Geodesic multi-modal mixup for robust fine-tuning. *arXiv preprint arXiv:2203.03897*.

Narges Tabari, Piyusha Biswas, Bhanu Praneeth, Armin Seyeditabari, Mirsad Hadzikadic, and Wlodek Zadrozny. 2018. Causality analysis of Twitter sentiments and stock market returns. In *Proceedings of the First Workshop on Economics and Natural Language Processing*, pages 11–19, Melbourne, Australia. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

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. Curran Associates, Inc.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR.

David Wilmot and Frank Keller. 2021. Memory and knowledge augmented language models for inferring salience in long-form stories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 851–865, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junran Wu, Ke Xu, Xueyuan Chen, Shangzhe Li, and Jichang Zhao. 2022. Price graphs: Utilizing the structural information of financial time series for stock prediction. *Information Sciences*, 588:405–424.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2560–2569.

Peng Xu, Xiatian Zhu, and David A. Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20.

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, Melbourne, Australia. Association for Computational Linguistics.

Chenggang Yan, Yunbin Tu, Xingzheng Wang, Yongbing Zhang, Xinhong Hao, Yongdong Zhang, and Qionghai Dai. 2020. Stat: Spatial-temporal attention mechanism for video captioning. *IEEE Transactions on Multimedia*, 22(1):229–241.

Soyoung Yoon, Gyuwan Kim, and Kyumin Park. 2021. SSMix: Saliency-based span mixup for text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3225–3234, Online. Association for Computational Linguistics.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Xianbing Zhao, Yixin Chen, Sicen Liu, Xuan Zang, Yang Xiang, and Buzhou Tang. 2023. Tmmda: A new token mixup multimodal data augmentation for multimodal sentiment analysis. In *Proceedings of the ACM Web Conference 2023*, pages 1714–1722.

Jie Zheng, Andi Xia, Lin Shao, Tao Wan, and Zengchang Qin. 2019. Stock volatility prediction based on self-attention networks with social information. In *2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pages 1–7.

Yan Zhou, Qingkai Fang, and Yang Feng. 2023. CMOT: Cross-modal mixup via optimal transport for speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7873–7887, Toronto, Canada. Association for Computational Linguistics.

9. Language Resource References

Castro et al. 2019. *MUStARD Dataset*. Association for Computational Linguistics. PID http://dx.doi.org/10.18653/v1/P19-1455.

Li et al. 2020. *Earnings Conference Call Dataset*. Association for Computing Machinery. PID https://doi.org/10.1145/3340531.3412879.

Sawhney et al. 2021. *Merger and Acquisitions Calls Dataset*. Association for Computational Linguistics. PID http://dx.doi.org/10.18653/v1/2021.acl-long.526.

Zadeh et al. 2016. *CMU-MOSI Dataset*. IEEE Intelligent Systems. PID https://doi.org/10.48550/arXiv.1606.06259.