Incorporating Stock Market Signals for Twitter Stance Detection

Costanza Conforti[1], **Jakob Berndt**[2], **Mohammad Taher Pilehvar**[1,3],
Chryssi Giannitsarou[2], **Flavio Toxvaerd**[2], **Nigel Collier**[1]

[1] Language Technology Lab, University of Cambridge
[2] Faculty of Economics, University of Cambridge
[3] Tehran Institute for Advanced Studies, Khatam University, Iran

{cc918,jb2088}@cam.ac.uk

Abstract

Research in stance detection has so far focused on models which leverage purely textual input. In this paper, we investigate the integration of textual and financial signals for stance detection in the financial domain. Specifically, we propose a robust multi-task neural architecture that combines textual input with high-frequency intra-day time series from stock market prices. Moreover, we extend WT–WT, an existing stance detection dataset which collects tweets discussing Mergers and Acquisitions operations, with the relevant financial signal. Importantly, the obtained dataset aligns with STANDER, an existing news stance detection dataset, thus resulting in a unique multimodal, multi-genre stance detection resource. We show experimentally and through detailed result analysis that our stance detection system benefits from financial information, and achieves state-of-the-art results on the WT–WT dataset: this demonstrates that the combination of multiple input signals is effective for cross-target stance detection, and opens interesting research directions for future work.

1 Introduction

Stance detection (SD) is the task of automatically classifying the writer's opinion expressed in a text towards a particular target (Küçük and Can, 2020). Starting from Mohammad et al. (2016)'s seminal work, research on Twitter SD gained increasing popularity (Ghosh et al., 2019), embracing new topics (Derczynski et al., 2017; Aker et al., 2017a; Conforti et al., 2020b) and languages (Gorrell et al., 2019; Vamvas and Sennrich, 2020a; Zotova et al., 2020). In recent years, research on SD has mainly focused on cross-target generalization, in which an SD system is tested on targets unseen during training (Xu et al., 2018). Cross-target generalization constitutes one of the biggest challenges in Twitter SD (AlDayel and Magdy, 2021): in this context, researchers investigated a wide range of techniques, including adversarial training (Wang et al., 2020; Allaway et al.), cross-lingual transfer (Mohtarami et al., 2019), knowledge transfer using semantic and emotion lexicons (Zhang et al., 2020), weak supervision through synthetic samples (Conforti et al., 2021b; Li and Caragea, 2021), and various types of cross-domain transfer (Schiller et al., 2021; Hardalov et al., 2021a).

In this paper, we study multimodality as a means to enhance cross-target generalization in Twitter SD. Multimodal Machine Learning studies the integration and modeling of multiple modalities (Elliott et al., 2016), where a *modality* refers to *the way in which something happens* (Baltrusaitis et al., 2019). Our contributions are as follows:

1. We study multimodal learning for Twitter SD. Despite being an established research area in NLP (Elliott et al., 2016), SD in a multimodal context is still understudied.

2. We extend WT–WT, an SD dataset which collects English tweets discussing four Mergers and Acquisitions operations (M&As or *mergers*, Conforti et al. (2020b)), with high frequency intra-day stock market data for the involved companies, which we release for future research[1]. We note that the union of our financial signal with WT–WT and with STANDER, an SD corpus collecting news articles discussing the same mergers (Conforti et al., 2020a), will constitute the first multi-genre, multi-modal parallel resource for SD and, more generally, one of the very few of this kind in NLP.

3. We propose SDTF (**S**tance **D**etection with **T**exual and **F**inancial signals), a novel multi-task, multimodal architecture for Twitter SD, which integrates textual and financial signals.

[1] https://github.com/cambridge-wtwt/acl2022-wtwt-stocks

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4074 - 4091

4. Finally, we show experimentally that SDTF benefits from the information encoded in the financial signal, achieving state-of-the-art results on the WT–WT dataset; the integration of multiple input signals thus constitutes a promising research direction to tackle cross-target generalization for SD.

2 Problem Formulation

We study SD in the financial domain and consider tweets discussing M&A operations, i.e. financial transactions in which the ownership of a company (the *target*) is transferred to another company (the *buyer*, Bruner and Perella (2004)). An M&A process usually comprises many stages, ranging from informal talks between the companies' boards to acquisition planning, negotiations, and external approvals, up to the closing of the deal (or its rejection, e.g. by antitrust bodies). M&As account for billions of dollars of investment globally and have been widely studied under many aspects (Gomes and Maldonado, 2020). They are well known in NLP (Lefever and Hoste, 2016; Yang et al., 2020; Conforti et al., 2020a,b) and constitute an important application in other AI fields, with a strong focus on automatic prediction of the M&A outcome (Yan et al., 2016; Jetley and Ji, 2010; Moriarty et al., 2019; Venuti, 2021).

In our task, a model receives a tweet and a target merger, and has to predict the stance expressed by the tweet's author with respect to the likelihood of the merger to succeed:

- Target. *Company A will merge with company B*
- Tweet. *Federal judge rejects A's bid to buy B!!!*
- Stance. *Refute*

All existing models for financial SD only leverage the tweet's text as input (Conforti et al., 2020b; Liang et al., 2021; Li and Caragea, 2021). However, a user tweeting at a particular time is immersed into a *context* which shapes their view of the world: their opinion about an M&A's outcome will be influenced by how the involved companies are perceived.

In this paper, we use a variation of the stock market prices from the n days prior to a tweet's posting as a means to provide a model with such context. According to the Efficient Market Hypothesis (Fama, 1970), stock market prices reflect all publicly known information. Even though the Efficient Market Hypothesis is controversial (Malkiel, 2003), stock market prices still reflect a considerable amount of publicly known information. Therefore, we argue that they can be used as a proxy for the available knowledge about the merger at a given time.

The relationship between rumors about an M&A operation and their effect on the involved companies' stocks is mutual and has been widely studied in finance (Ma and Zhang, 2016; Betton et al., 2018; Jia et al., 2020; Gorman et al., 2021; Davis et al., 2021), but never investigated in NLP. To our knowledge, the integration of textual and financial data signals has been studied for financial forecasting (Schumaker and Chen, 2009; Hu et al., 2018; Sawhney et al., 2020a,b, 2021c; Ni et al., 2021), but has yet to be investigated for SD.

3 Background

3.1 Twitter SD

Traditionally, research on SD has focused on user-generated data, such as blogs and commenting sections on websites (Skeppstedt et al., 2017; Hercig et al., 2017), apps (Vamvas and Sennrich, 2020b), online debate forums (Somasundaran and Wiebe, 2009), Facebook posts (Klenner et al., 2017) and, above all, Twitter. Since Mohammad et al. (2016)'s seminal work, Twitter has been used as a data source for collecting corpora covering a wide range of domains, from US politics (Mohammad et al., 2017; Inkpen et al., 2017) to mental health (Aker et al., 2017b), breaking news events (Zubiaga et al., 2016; Gorrell et al., 2019), finance (Conforti et al., 2020b), and the COVID pandemic (Hossain et al., 2020; Glandt et al., 2021).

SD has been studied both as a stand-alone, isolated task, and integrated as a sub-component of more complex NLP pipelines (Hardalov et al., 2021b). Starting from the pioneering work by Vlachos and Riedel (2014), SD has been identified as a key step in fake news detection (Lillie and Middelboe, 2019) and automated fact-checking (Popat et al., 2017; Thorne and Vlachos, 2018; Baly et al., 2018).

3.2 Multimodal SD

Multimodal learning has proven successful for many NLP tasks (Tsai et al., 2019; Zadeh et al., 2020), including grounding (Beinborn et al., 2018), visual question answering (Ben-Younes et al., 2017; Yu et al., 2018), sentiment analysis (Rahman et al., 2020), and humor detection (Hasan et al., 2019).

To the best of our knowledge, only one

M&A	Buyer	Target	Outcome
CVS_AET	CvsHealth	Aetna	yes
CI_ESRX	Cigna	ExprsScripts	yes
ANTM_CI	Anthem	Cigna	no
AET_HUM	Aetna	Humana	no

Table 1: Healthcare M&As in WT–WT. AET and CI appear both as buyers and as targets.

	CSV AET	CI ESRX	ANTM CI	AET HUM
support	2,469	773	970	1,038
refute	518	253	1,969	1,106
comment	5,520	947	3,098	2,804
unrelated	3,115	554	5,007	2,949
total	11,622	2,527	11,622	7,897

Table 2: Label distribution across M&As in the WT–WT corpus (total: 33,090 tweets).

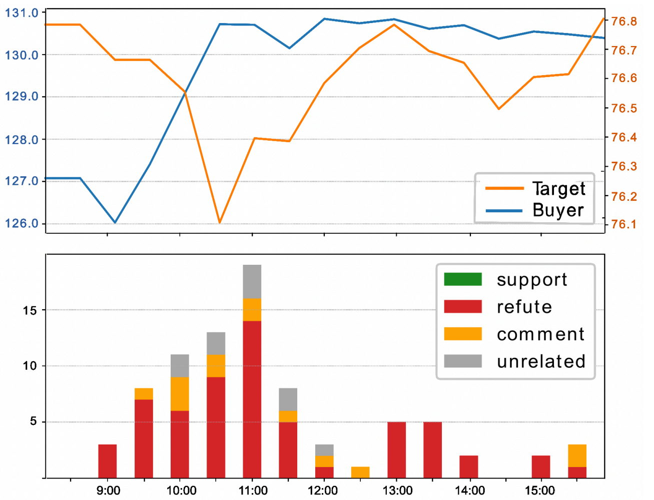

Figure 1: Stock prices of ANTM (buyer) and CI (target) and tweets distribution on the day of the official antitrust complaint to the Department of Justice (21.07.2013).

dataset exists for multimodal SD, MULTISTANCECAT (Taulé et al., 2018; Segura-Bedmar, 2018), released for IberEval2018[2]. MULTISTANCECAT collects 11,398 tweets in Spanish and Catalan discussing the Catalan 2017 Independence referendum: according to Taulé et al. (2018), the corpus is multimodal because it contains, along with the tweets' text, contextual information and up to 10 images downloaded from the authors' timeline. We note that, unfortuntately, almost all research building on MULTISTANCECAT considered only the provided textual features, thus ignoring its multimodal component. As mentioned in Taulé et al. (2018, p. 157), only 1 out of the 4 teams participating in the task integrated images into their model, by training a CNN on Spanish and Catalan flags (with the underlying intuition that using them would hint to the user's stance with respect to the topic of Catalan independence)[3]. Interestingly, no positive impact was observed on SD results when including such multimodal signals.

Our work differs in a number of respects: (1) the size of our corpus is considerably larger, thus allowing for more robust training; (2) we do not consider visual signals, such as images, but – consistently with WT–WT's domain – financial time-series signals from stock market prices; and (3) most notably, MULTISTANCECAT's multimodal signal consists of a maximum of 10 images taken from the user's timeline: therefore, the images might not be related to the tweet, might have been posted at a very different timestamp, or might be the same for multiple tweets published by the same author. In contrast, our financial signal is specific to each tweet and is perfectly aligned with its time of posting.

3.3 Finance and NLP

In recent years, there has been an increasing interest in research at the intersection between finance and NLP (Hahn et al., 2018; El-Haj et al., 2018), with a rich stream of work focusing on financial textual analysis (Lang and Stice-Lawrence, 2015; Loughran and McDonald, 2016), sentiment analysis (Giachanou and Crestani, 2016; Chan and Chong, 2017; Krishnamoorthy, 2018), stance detection (Conforti et al., 2020b,a, 2021a), volatility prediction (Rekabsaz et al., 2017; Kolchyna et al., 2015) and, above all, financial forecasting (Qasem et al., 2015; Ranco et al., 2015; Pagolu et al., 2016; Pimprikar et al., 2017; Oliveira et al., 2017).

3.4 Multimodality in Financial Forecasting

While multimodality has not been investigated for financial SD, it constitutes a very active research direction in financial forecasting, i.e. the task of predicting a business' future financial performance (Abu-Mostafa and Atiya, 1996).

Given the importance of psychological and behaviorial elements on stock-price movements (Malkiel, 2003), researchers in economics

[2] http://www.autoritas.net/MultiStanceCat-IberEval2018/

[3] The team did not submit working notes describing their system; therefore, we refer to the model's overview provided in the general task paper (Taulé et al., 2018).

have started to explore models which leverage features beyond simple numerical values (Nikou et al., 2019; Liu and Chen, 2019). In this context, a stream of work analyzed the integration of historical price data with social media texts (Sawhney et al., 2020a) and other audio or textual features (Zhao et al., 2019; Qin and Yang, 2019; Sawhney et al., 2021b; Lee and Yoo, 2020; Sawhney et al., 2021b,a; Das et al., 2021; Chen and Huang, 2021).

4 Extending the WT–WT Dataset

Text Signal. As our text signal, we use Will-They-Won't-They (Conforti et al., 2020b, WT–WT)[4], which collects English tweets discussing four M&As between US companies (Table 1). WT–WT is expert-annotated for stance with respect to the likelihood of the merger happening according to the opinion expressed in the text, following a four-class classification schema: *support, refute, comment* and *unrelated* (i.e. the tweet does not discuss the merger). Below, we report one example for each of the considered labels (targets in squared brackets):

- Support [CVS_AET] *CVS, Aetna $69B merger wins DOJ approval <URL>*
- Refute [ANTM_CI] *Big-name lawmakers want to block Aetna-Humana and Anthem-Cigna!*
- Comment [ANTM_CI] *Anthem-Cigna deal would create 'Big 3': If the deal is approved*
- Unrelated [CVS_AET] *Urge Your Legislators to Oppose CVS and Walmart Takeover of Medical Care Delivery!!! <URL>#MSSNY*

Financial Signal. For the four healthcare M&As in WT–WT[5], we obtain historical prices in 30-min intervals for the involved stocks. The financial data has been bought from FirstRate Data LLC[6] (~700MB) at market price.

Each entry in the data has the following fields: *DateTime, Open, High, Low, Close, Volume. DateTime* is in US Eastern Time, in the format `YY-MM-DD h:m:s`. Only minutes with trading volume are included: times with zero volume, such as during weekends or holidays, are omitted. Prices are adjusted for dividends and splits[7]. We used Python's `datetime` library to align Twitter time values (UTC) with the financial signal (EST, New York Stock Exchange)[8]

Note that price variations in 30-minutes intervals are considerably more granular than the financial signal used in NLP work, which is mostly limited to daily data (Sawhney et al., 2020a). Such granularity is necessary when monitoring tweets, which are highly reactive to real-time, on-topic information from the outside world (ALRashdi and O'Keefe, 2019).

Analysis. Figure 1 shows an example of the integration of the two signals. On the day the antitrust complaint was made to the Department of Justice regarding the M&A operation, ANTM's price increased while CI's decreased. Such movements testify that the event changed the world's view: people believe that the merger is less likely to happen, and this is reflected by their investment decisions. The direction of the price variation reflects standard M&A theory (Bruner and Perella, 2004): the buyer will not buy the target's shares at a premium, thus the owners of target's stocks will not profit from the acquisition.

The price variation is useful for classifying a tweet on that day, as it implies that the likelihood of a *refute* label is higher. This is reflected in the tweet distribution in the lower part of the Figure: the distribution of tweets on that day shows that most of them were indeed *refuting*. We report one more example in Appendix A.

5 Models

As shown in Figure 2, our multitask SDTF model is composed of a *textual*, a *financial* and a *multimodal* component.

5.1 Text Encoder

Following previous work in SD (Hardalov et al., 2021a), we obtain a vector representation $h_{text} \in \mathbb{R}^d$ for the textual input by averaging the token-level hidden states from the last layer of a large transformer (in our case, BerTweet (Nguyen et al.,

[4] WT–WT can be downloaded, upon signing a data sharing agreement, from its GitHub repository https://github.com/cambridge-wtwt/acl2020-wtwt-tweets

[5] Note that this aligns with the targets collected in STANDER, a news SD corpus (Conforti et al., 2020a).

[6] https://firstratedata.com/

[7] https://firstratedata.com/about/price_adjustment

[8] The timestamps of posting of each tweet in the WT–WT dataset can be shared in accordance with the terms of use outlined by Twitter https://developer.twitter.com/en/developer-terms/agreement-and-policy. No private information (such as username of the tweet's author and similar) is shared.

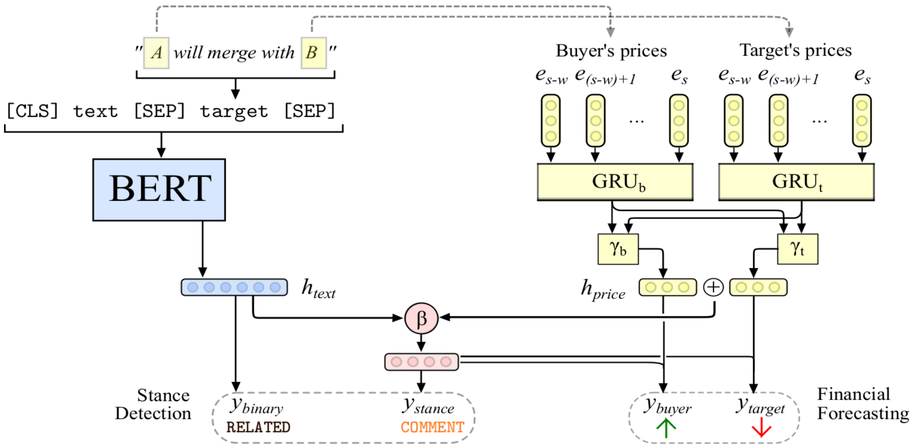

Figure 2: Overview of the proposed multi-task SDTF architecture. Price embeddings are not shown. Right, middle, and left components represent resp. textual, blended and financial signals. γ is a multi-head attention mechanism, and β is a bilinear transformation (Subsection 5.3).

2020)). The input text is provided as:

`[CLS] tweettext [SEP] target [SEP]`

where `target` consists of the string: B (b, t_b) *will merge with* T (t, t_t), where B, b, and t_b, are the buyer's name, acronym and Twitter username[9] (same for the target company).

5.2 Price Encoder

Input. For each tweet posted at time s, we consider a window of w days in the past. At each timestep i, in $\{s-w, s-w+1, ..., s\}$, we consider two price vectors $p_i^b, p_i^t \in \mathbb{R}^{12}$ which consist of:

$$\begin{aligned} p_i^b &= p_{i1}^b \oplus p_{i2}^b \oplus p_{i3}^b \\ &= [o^b, c^b, h^b, l^b] \oplus [o^m, c^m, h^m, l^m] \\ &\quad \oplus [v^b, r^b, \frac{c^b}{c^m} \frac{r^b}{c^m}] \end{aligned} \tag{1}$$

where o, c, h, l and v are resp. the opening, closing, highest, lowest price and volume of transactions at time i for the buyer's stock (superscript b) or for the overall market index (superscript m); finally, r is the return at time i and is defined as $(c_i^b - c_{i-1}^b)/c_{i-1}^b$ (Law (2018), same for the target).

Price Embeddings. We obtain a vector representation e_b^i for each time point i by concatenating:

$$p_i^b \oplus e_{i1}^b \oplus e_{i2}^b \tag{2}$$

where e_{i1}^b and e_{i2}^b are the time embeddings for p_{i1}^b and p_{i2}^b (same for the target). We use Time2Vec (Kazemi et al., 2019) for time embeddings, and we jointly learn embeddings for the buyer and the target.

Price Encoder. As in Du and Tanaka-Ishii (2020) and Kostkova et al. (2017), we use a Gated Recurrent Unit (Cho et al., 2014, GRU) to encode the price variations over time. We implement two separate GRU_b and GRU_t for the buyer and the target. At time i, the GRU_b's output consists of:

$$h_i = GRU_b(e_b^i, h_{i-1}) \quad s-w \leq i \leq s \tag{3}$$

To model the inter-dependencies between the two stocks, we use multi-head attention mechanism (Vaswani et al., 2017) which, in our experiments, proved to be more effective for SD than the "classic" temporal attention used in financial forecasting (Feng et al., 2019). In practice, we obtain a unified price vector representation h_{price} as:

$$h_b = \gamma_b(H_t, H_b) \tag{4}$$

$$h_t = \gamma_t(H_b, H_t) \tag{5}$$

$$h_{price} = h_b \oplus h_t \tag{6}$$

where γ_b and H_b (resp. γ_t and H_t) are the buyer's (and target's) multi-head attention mechanism and the matrix consisting of GRU_b's (resp. GRU_t's) outputs.

5.3 Blending Multimodal Signals

Signals from different modalities encode complementary information (Schumaker and Chen, 2009): we avoid simple concatenation (Li et al., 2016),

[9] For example, *Anthem (ANTM, AnthemInc)*. This is in principle the same as in (Liang et al., 2021), with two differences: we add the companies' official Twitter usernames and, similarly to other SD works (Hardalov et al., 2021a), we consider first the input text, and then the target.

which would treat such signals equally, and implement a bilinear transformation to integrate the tweet's encoded representation with the historical prices of the involved companies (Sawhney et al., 2020a). Given the price and the text vector representations $h_{price} \in \mathbb{R}^p$ and $h_{text} \in \mathbb{R}^d$, we obtain a combined vector representation $h \in \mathbb{R}^w$ as:

$$h = relu(h_{text}^T W h_{price} + b) \tag{7}$$

where $W \in \mathbb{R}^{w \times d \times p}$ and $b \in \mathbb{R}^w$ are the learned weight matrix and bias.

5.4 Multi-Task Training

We jointly train our model to learn two sets of tasks: SD and financial forecasting (FF).

Stance Detection. We expect the financial signal to be relevant only in the case of *related* stance labels (i.e. *support, refute, comment*). In order to assist the model in differentiating between those two macro-classes, we predict a binary label *related/unrelated* along with the stance label y_{stance}:

$$y_{stance} = \text{softmax}(h) \quad y_{binary} = \sigma(h_{text}) \tag{8}$$

Financial Forecasting. As it has been previously studied in finance, rumors about a merger can affect the stock prices of the involved companies (Jia et al., 2020; Davis et al., 2021). To encourage our model to learn such influence, we also add two binary financial-related outputs, in which we predict the stock movement of the two companies:

$$y_{buyer} = \sigma(h_{buyer}) \tag{9}$$

$$y_{target} = \sigma(h_{target}) \tag{10}$$

where h_{buyer} (resp. h_{target}) is the concatenation of the last output vector of GRU_b and h, and y_{buyer} (resp. y_{target}) $\in \{\uparrow, \downarrow\}$ (i.e., stock closing price for the considered company will resp. move up, or fall). The final loss is:

$$\mathcal{L} = \mathcal{L}_{stance} + 0.5\mathcal{L}_{binary} + 0.2\mathcal{L}_{buyer} + 0.2\mathcal{L}_{target} \tag{11}$$

For $\mathcal{L}_{stance}$ we use categorical cross-entropy loss, while $\mathcal{L}_{binary}, \mathcal{L}_{buyer}$ and $\mathcal{L}_{target}$ use binary cross-entropy loss function. The weights of the last three loss components were empirically set in an initial pilot.

6 Experimental Setting

Preprocessing. We perform minimal preprocessing on the textual signal. Concerning the financial signal, we consider a window of 30 timepoints in the past, and price variations every 30 minutes: depending on the tweet's posting time, this accounts for the previous ~2.5 days[10].

For FF, we predict ups or downs in the considered company's closing price 2 hours after the tweet[11] (see Appendix B.1 for details).

Training Setup and Evaluation. Details on the training setup and (hyper-)parameter settings are reported in Appendix B.2 for replication. Following Hanselowski et al. (2018); Conforti et al. (2020b), we consider macro-averaged precision, recall and F_1 score. To account for performance fluctuations (Reimers and Gurevych, 2017), we average three runs for each model (standard deviation is reported in Appendix B.2).

Baselines. We consider six published baseline models, including the four best models of Conforti et al. (2020b):

- *SVM*, a linear-kernel SVM leveraging bag of ngrams (over words and characters) features, similar as in Mohammad et al. (2017);
- *CrossNet*, a cross-target SD model (Xu et al., 2018) consisting of a bidirectional conditional encoding model over LSTMs, augmented with self-attention and two dense layers;
- *SiamNet*, a siamese network similar to Santosh et al. (2019), which is based on a BiLSTM followed by a self-attention layer;
- *HAN*, a Hierarchical Attention Network as in (Sun et al., 2018)) which uses two levels of attention to leverage the tweet representation along with linguistic information (sentiment, dependency and argument);

and two further baselines from Liang et al. (2021):

- *BERT*, a strong vanilla BERT-based model fine-tuned on WT–WT;
- *TPDG*, a sophisticated network based on a target-adaptive pragmatics dependency graph.

[10] During night or holidays, price entries are usually not available. Tweets published outside of the market's opening hours (9:30am–4pm EST during workdays) are thus associated with the most recent available financial signal.

[11] Or, for tweets posted at night or during holidays, the first available closing price in the future.

	CVS_AET	CI_ESRX	ANTM_CI	AET_HUM	$avgF_1$	avg_wF_1	*sup*	*ref*	*com*	*unr*
SVM$^\natural$	51.0	51.0	65.7	65.0	58.1	58.5	54.5	43.9	41.2	**88.4**
CrossNet$^\natural$	59.1	54.5	65.1	62.3	60.2	61.1	63.8	48.9	50.5	75.8
SiamNet$^\natural$	58.3	54.4	68.7	67.7	62.2	63.1	67.0	48.0	52.5	78.3
HAN$^\natural$	56.4	57.3	66.0	67.3	61.7	61.7	67.6	52.0	55.2	69.1
BERT$^\flat$	56.0	60.5	67.1	67.3	62.7	62.8	65.4	56.1	58.0	70.1
TPDG$^\flat$	66.8	65.6	74.2	73.1	69.8	70.7	69.7	64.9	69.8	76.9
BerTweet	71.7	70.4	70.8	69.6	70.6	70.4	70.0	66.2	70.2	75.9
SDTF (ST)	71.5	**73.7**	74.3	75.5	73.7	73.8	**75.4**	68.2	72.7	79.6
SDTF (MT)										
+FF	72.3	73.2	76.0	75.7	74.3	74.0	74.8	67.2	73.7	81.6
+Binary	70.4	73.4	**77.1**	74.8	73.9	73.4	73.2	67.7	73.5	78.9
+FF+Binary	**72.9**	72.7	77.0	**78.1**	**75.2**	**74.9**	75.2	**68.6**	**74.3**	82.7

Table 3: Results on the WT–WT dataset. Macro F_1 are obtained by testing on a target M&A while training on the other three. $avgF_1$ and avg_wF_1 are the unweighted and weighted (by operations size) avg over targets. On the right, average per-label accuracy. $\natural$ and $\flat$ results are retrieved resp. from Conforti et al. (2020b) and Liang et al. (2021). MT is the complete multitask model in Figure 2, ST refers to a single-task model trained for SD only.

Finally, we also consider *BerTweet*, a model relying on textual signal only; it is a BerTweet model (Nguyen et al., 2020) fine-tuned on WT–WT.

7 Results and Discussion

Table 3 shows our experimental results. We observe that using BerTweet as main text encoder alone achieves considerable gains in performance with respect to all stance labels considering all baselines, including the strong vanilla BERT baseline.

This is unsurprising, given the peculiarities of Twitter language (Hu et al., 2013) which are captured by BerTweet.

Adding the financial signal. Adding our financial component proves to be effective over all considered targets, with improvements in F_1 scores up to +5.8 (AET_HUM).

Single-label performance seems to suggest that price variations encode very useful information for all labels, resulting in notable improvements not only on the *unrelated* (+3.7), but also on the *refute* and *support* samples (resp. +2.1 and +5.4 in accuracy): this is important because those labels, apart from being the minority classes, arguably constitute the most relevant information for downstream tasks (Scarton et al., 2020).

Adding Multi-Task Objectives and Ablation Experiments. Results of ablation experiments (Table 3) show that including the financial forecast (+FF) task alone brings moderate improvements in performance, while considering binary SD (+Binary) alone moderately degrades it: their combination, however, achieves the best results over three of the four mergers.

		CSV AET	CI ESRX	ANTM CI	AET HUM	avg.
+FF	buyer FF	51.3	49.6	48.9	51.4	50.3
	target FF	41.9	52.9	52.5	53.8	50.3
+Bin	SD bin	85.4	88.5	93.0	85.7	88.1
+FF +Bin	SD bin	86.3	89.8	92.6	90.6	**89.8**
	buyer FF	48.7	53.6	52.1	49.2	**50.9**
	target FF	52.0	51.5	49.8	50.2	**50.9**

Table 4: Per-merger performance (binary accuracy) of the SDTF multitask models on the ancillary tasks. *+FF*: financial forecasting; *+Bin*: binary SD.

Interestigly, jointly modeling FF and binary SD seems to be beneficial not only for SD: as shown in Table 4, best results on both ancillary tasks are obtained in the multitask setting. Binary SD performance is very satisfactory over all mergers, with a correlation with M&As with a higher proportion of *unrelated* samples.

Moving to the other ancillary tasks, FF results are encouraging[12], even if we considered a considerably shorter time window of historical pricing than architectures specifically designed for FF (Dumas et al., 2009; Kim et al., 2019; Ho et al., 2021). This suggest that the learned multimodal textual and financial vectors constitute an informative input for the FF predictors.

Single-Label Performance. An analysis of single-label performance (Table 3) shows that models including the financial component, with or without ancillary tasks, achieve best performance on all related labels.

[12]Consider for example a strong neural model such as Selvin et al. (2017), reported in (Sawhney et al., 2020a).

SDTF (MT)	*sup*	*ref*	*com*	*unr*	avg. F_1
text only	70.8	66.6	68.7	74.9	70.4
financial only	0.0	2.2	27.2	46.7	21.1
text+financial	75.2	68.6	74.3	82.7	75.2

Table 5: Ablation experiments with multi-task SDTF when "silencing" the textual or financial signal (per-label average accuracy and average F_1 score over mergers); *text+financial* corresponds to the complete SDTF model in Table 3.

	Precision	Recall	F_1 score
BerTweet (frozen)	60.34	58.04	56.69
" (frozen:9)	73.36	74.66	73.63
" (train all)	72.18	71.02	70.62
SDTF (frozen)	67.04	66.95	63.08
" (frozen:9)	73.83	74.96	74.15
" (train all)	74.85	76.39	75.19

Table 6: Average model performance over targets of our multitask multimodal system, when partially freezing TweeBert layers.

Interestingly, however, best performance overall for the *unrelated* samples is obtained with the simplest of the considered models, a strong SVM over character- and word-ngrams similar to (Mohammad et al., 2017). A similar situation, in which a model leveraging simple lexical features achieved best results on the *unrelated* samples, was already observed not only for WT–WT (Conforti et al., 2020b), but also for other SD datasets, such as FNC-1 (Pomerleau and Rao, 2017; Hanselowski et al., 2019).

We note that, in both datasets, *related-unrelated* vs. *support/comment/refute* classifications can be seen as constituting two different tasks: the former is more similar to topic detection, where even surface-level methods can do well, whereas the latter is an inference task which requires deeper semantic knowledge (Conforti et al., 2018)[13].

The analysis of the confusion matrices (reported in detail in Appendix B.2) shows that most errors concern *support* or *refute* samples which were misclassified as *comment*: as already observed in Conforti et al. (2020b), the difference between a *comment* and a stance-bearing label such as *support* (or *refute*) depends on argumentative nuances in the tweet, which are sometimes subjective and ultimately depends on the annotator's preferences. A number of *comment-unrelated* misclassifications are also present, especially for M&As with a high number of *unrelated* samples (such as CVS_AET and ANTM_CI).

Performance When "Silencing" Different Signals. In order to estimate the relative importance of the two signals considered in the SDTF model, we consider a scenario in which we silence one of the two signals: for the textual signal, this corresponds to replacing the target and the tweet's text with two empty strings (i.e., `[CLS] [SEP] [SEP]` as input to the right component in Figure 2); for the financial signal, we input two empty price vectors for the considered companies (i.e. the left components in Figure 2).

Results of such ablation experiments (Table 5) show that, as expected, the textual signal provides the biggest contribution for SD, and the financial signal alone is not sufficient at all to perform SD. Blending together both signals, however, provides the most informative input to the model: a consistent drop in performance over all labels, including *unrelated*, is observed with models exposed to empty price vectors.

Robustness Over Parameters Freezing. Moreover, we investigate the model robustness over freezing BerTweet[14]: we consider two scenarios, in which we freeze the complete weights or BerTweet, or all but its last three layers (Wang et al. (2019), see Appendix B.2 for details on number of parameters for the different settings).

As expected (Mosbach et al., 2020), performance degrades with fewer layers trained (Table 6), with the exception of the BerTweet architecture when freezing all but its last three layers. Notably, our multitask SDTF model is more robust over parameter freezing than the vanilla BerTweet, achieving higher performance over all considered metrics: this suggests that, when less powerful textual encoders are provided, the presence of the financial signal supports SD classification.

Adding Synthetic Data. As mentioned in the Introduction, a recent stream of work investigates the usage of synthetically generated data to compensate for data scarcity in Twitter SD. In particular, Li and Caragea (2021) used Auxiliary Sentence based Data Augmentation (ASDA), a conditional

[13] We note that, in a practical scenario, it might make sense to first apply a simple lexicon-based method for filtering out *unrelated* samples, and then to adopt a more sophisticated approach for the second step, as proposed for example by Masood and Aker (2018).

[14] This is important, because the number of trainable parameters correlates with CO_2 emission (Strubell et al., 2019).

	CSV AET	CI ESRX	ANTM CI	AET HUM	avg.
ASDA$^\sharp$	**76.4**	75.4	74.5	79.0	76.5
SDTF	72.9	72.7	77.0	78.1	75.2
ASDA + SDTF	74.6	**75.9**	**77.8**	**79.7**	**77.0**

Table 7: Per-merger performance (F_1 score) when including synthetic training data. $\sharp$ results refer to the ASDA$_{\text{WT–WT}}$ model and are retrieved from Li and Caragea (2021); SDTF indicates our multi-task model.

data augmentation method, to double the size of SD datasets, achieving state-of-the-art results on WT–WT with a model trained on the union of gold and synthetic samples.

In a last set of experiments, we investigate the impact of adding such synthetically generated examples to an SDTF model. As synthetic samples aren't associated to any price vectors from the stock market, we proceed as follows: we first fine tune a *BerTweet* model on ASDA$_{\text{WT–WT}}$, which we obtain from the ASDA paper's authors; then, we use such model's weights to initialize the textual encoder of an SDTF multitask model (the left components in Figure 2), which we finally train on the gold WT–WT as described in Section 5.

Results in Table 7 show that models trained on ASDA$_{\text{WT–WT}}$ (gold and synthetic samples) achieve better results than SDTF trained on gold data alone. Including synthetic signal from ASDA$_{\text{WT–WT}}$ seems to be effective for all considered training settings: even using a simple pretraining strategy as described above allows an SDTF model to capture useful textual features from the synthetic samples, which are retained over the finetuning stage and allow for better cross-target generalization.

Our finetuned model (ASDA+SDTF in Table 7) reaches state-of-the-art results on the WT–WT dataset and best results over three of the four considered mergers, with gains in F_1 scores ranging from $+1.4$ (ANTM_CI) to $+3.2$ (CI_ESRX).

8 Conclusions

In this paper, we studied the well-established task of Twitter SD in a multitask scenario, focusing on the financial domain. We proposed SDTF, a novel model which integrates two modalities, text and financial time series data. We extended WT–WT, a large dataset for financial SD, with financial signals from stock market prices. Our detailed analysis of models' results demonstrated that financial SD on tweets benefits from such signals: models which include textual and financial features showed better cross-target generalization capabilities, and obtained better results on all stance labels. Finally, we proposed a simple but effective setting to leverage useful signals encoded in synthetic samples, reaching state-of-the-art results on WT–WT.

We release the financial signal collected to complement WT–WT: together with the STANDER corpus of news SD, which discusses the same mergers, it constitutes an invaluable and unique resource to foster research on multi-modal, multi-genre SD, and to model the integration and mutual influences between stock market variations, tweets, and authoritative news sources.

Ethics and Broader Impact

Data Collection. Daily financial data is publicly available and can be freely downloaded (e.g. through Yahoo Finance[15]). However, granular financial data needs to be purchased. We bought the historical financial data from FirstRate Data LLC[16], who source their data directly from major exchanges. We tested all signals for consistency and completeness, and found that it reflects the actual trading in the stocks.

Presence of Bias. As textual input, we used WT–WT, a publicly available dataset which we obtained from the authors after signing a data sharing agreement (Academic Free License). Given that many NLP tasks are somehow subjective (Poesio et al., 2019), and the choice of annotators might reinforce the emergency of bias (Waseem, 2016; Sap et al., 2019; Geva et al., 2019) we note that WT–WT might contain annotation bias, which could be amplified by our models (Shah et al., 2020; Waseem et al., 2021). Moreover, the BerTweet model we are using as main text encoder might encode biases due to the data it was trained on (Bender et al., 2021). We observe, however, that both elements are beyond our control.

Data Sharing. In accordance with FirstRate Data, we release the relevant portion of the data under Academic Free License at the link: `https://github.com/cambridge-wtwt/acl2022-wtwt-stocks`. We are aware of the many ethical issues surrounding social media research (Hovy and Spruit, 2016). Virtually all models trained on social media data are dual-use (Benton et al., 2017): in order to avoid

[15] `https://uk.finance.yahoo.com/`
[16] `https://firstratedata.com/`

potential misuse, we will share our financial signals, which is complementary to WT–WT, only upon signing a data sharing agreement restricting the data usage to research only.

Environmental Factors. We are conscious that training transformers such as BerTweet produces large quantity of CO_2 emissions (Strubell et al., 2019; Henderson et al., 2020). We observe that, in our case, we are not training such models from scratch, thus considerably limiting the training time. Moreover, we also experimented with (partially) frozen transformers (Lee et al., 2019; Sajjad et al., 2020; Mosbach et al., 2020), which in turn require less parameters to be optimized.

Acknowledgments

We thank the anonymous reviewers of this paper for their efforts and for the constructive comments and suggestions. We gratefully acknowledge funding from the Keynes Fund, University of Cambridge (grant no. JHOQ). CC is grateful to NERC DREAM CDT (grant no. 1945246) for partially funding this work. CG and FT are thankful to the Cambridge Endowment for Research in Finance (CERF).

References

Yaser S Abu-Mostafa and Amir F Atiya. 1996. Introduction to financial forecasting. *Applied intelligence*, 6(3):205–213.

Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. 2017a. Simple open stance classification for rumour analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 31–39.

Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, Anna Kolliakou, Rob Procter, and Maria Liakata. 2017b. Stance classification in out-of-domain rumours: A case study around mental health disorders. In *Social Informatics - 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II*, volume 10540 of *Lecture Notes in Computer Science*, pages 53–64. Springer.

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Emily Allaway, Malavika Srikanth, and Kathleen McKeown. Adversarial learning for zero-shot stance detection on social media. In *To appear in: Proceedings of NAACL 2021*.

Reem ALRashdi and Simon O'Keefe. 2019. Deep learning and word embeddings for tweet classification for crisis response. *arXiv preprint arXiv:1903.11024*.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

Ramy Baly, Mitra Mohtarami, James R. Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 21–27. Association for Computational Linguistics.

Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. 2018. Multimodal grounding for language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2325–2339, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620.

Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors. 2018. *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, EthNLP@EACL, Valencia, Spain, April 4, 2017*, pages 94–102. Association for Computational Linguistics.

Sandra Betton, Frederick Davis, and Thomas Walker. 2018. Rumor rationales: The impact of message justification on article credibility. *International Review of Financial Analysis*, 58:271–287.

Robert F Bruner and Joseph R Perella. 2004. *Applied mergers and acquisitions*, volume 173. John Wiley & Sons.

Samuel WK Chan and Mickey WC Chong. 2017. Sentiment analysis in financial texts. *Decision Support Systems*, 94:53–64.

Yu-Fu Chen and Szu-Hao Huang. 2021. Sentiment-influenced trading system based on multimodal deep reinforcement learning. *Applied Soft Computing*, 112:107788.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL.

Costanza Conforti, Jakob Berndt, Marco Basaldella, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2021a. Adversarial training for news stance detection: Leveraging signals from a multi-genre corpus. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 1–7, Online. Association for Computational Linguistics.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020a. STANDER: an expert-annotated dataset for news stance detection and evidence retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 4086–4101. Association for Computational Linguistics.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020b. Will-they-won't-they: A very large dataset for stance detection on twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1715–1724. Association for Computational Linguistics.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2021b. Synthetic examples improve cross-target generalization: A study on stance detection on a Twitter corpus. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 181–187, Online. Association for Computational Linguistics.

Costanza Conforti, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Towards automatic fake news detection: Cross-level stance detection in news articles. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 40–49, Brussels, Belgium. Association for Computational Linguistics.

Sanjiv Das, Connor Goggins, John He, George Karypis, Sandeep Krishnamurthy, Mitali Mahajan, Nagpurnanand Prabhala, Dylan Slack, Rob van Dusen, Shenghua Yue, et al. 2021. Context, language modeling, and multimodal data in finance. *The Journal of Financial Data Science*.

Frederick Davis, Hamed Khadivar, and Thomas J Walker. 2021. Institutional trading in firms rumored to be takeover targets. *Journal of Corporate Finance*, 66:101797.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 69–76.

Xin Du and Kumiko Tanaka-Ishii. 2020. Stock embeddings acquired from news articles and price history, and an application to portfolio optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3353–3363. Association for Computational Linguistics.

Bernard Dumas, Alexander Kurshev, and Raman Uppal. 2009. Equilibrium portfolio strategies in the presence of sentiment risk and excess volatility. *The Journal of Finance*, 64(2):579–629.

Mahmoud El-Haj, Paul Rayson, and Andrew Moore. 2018. Proceedings of the first financial narrative processing workshop. In *Proceedings of the 11th Language Resources and Evaluation Conference, Miyazaki, Japan*.

Desmond Elliott, Douwe Kiela, and Angeliki Lazaridou. 2016. Multimodal learning and reasoning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, Berlin, Germany. Association for Computational Linguistics.

Eugene F. Fama. 1970. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. 2019. Enhancing stock movement prediction with adversarial training. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5843–5849. ijcai.org.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 1161–1166. Association for Computational Linguistics.

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance detection in web and social media: a comparative study. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 75–87. Springer.

Anastasia Giachanou and Fabio Crestani. 2016. Like it or not: A survey of twitter sentiment analysis methods. *ACM Comput. Surv.*, 49(2):28:1–28:41.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. 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 1596–1611, Online. Association for Computational Linguistics.

Armando Gomes and Wilfredo Maldonado. 2020. Mergers and acquisitions with conditional and unconditional offers. *Int. J. Game Theory*, 49(3):773–800.

Louise Gorman, Theo Lynn, Eleonora Monaco, Riccardo Palumbo, and Pierangelo Rosati. 2021. The effect of media coverage on target firms' trading activity and liquidity around domestic acquisition announcements: evidence from uk. *The European Journal of Finance*, pages 1–20.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of SemEval 2019*.

Udo Hahn, Véronique Hoste, and Ming-Feng Tsai. 2018. Proceedings of the first workshop on economics and natural language processing. In *The 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia*.

Andreas Hanselowski, Avinesh P. V. S., Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In (Bender et al., 2018), pages 1859–1874.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021a. Cross-domain label-adaptive stance detection. *CoRR*, abs/2104.07467.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021b. A survey on stance detection for mis- and disinformation identification. *CoRR*, abs/2103.00242.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *CoRR*, abs/2002.05651.

Tomáš Hercig, Peter Krejzl, Barbora Hourová, Josef Steinberger, and Ladislav Lenc. 2017. Detecting stance in czech news commentaries. In *Proceedings of the 17th ITAT: Slovenskoceský NLP workshop (SloNLP 2017)*, volume 1885, pages 176–180.

Tuan Ho, Ruby Brownen-Trinh, and Fangming Xu. 2021. The information content of target price forecasts: Evidence from mergers and acquisitions. *Journal of Business Finance & Accounting*, 48(5-6):1134–1171.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.

Yuheng Hu, Kartik Talamadupula, and Subbarao Kambhampati. 2013. Dude, srsly?: The surprisingly formal nature of twitter's language. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013*. The AAAI Press.

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 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 261–269. ACM.

Diana Inkpen, Xiaodan Zhu, and Parinaz Sobhani. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 551–557. Association for Computational Linguistics.

Gaurav Jetley and Xinyu Ji. 2010. The shrinking merger arbitrage spread: Reasons and implications. *Financial Analysts Journal*, 66(2):54–68.

Weishi Jia, Giulia Redigolo, Susan Shu, and Jingran Zhao. 2020. Can social media distort price discovery? evidence from merger rumors. *Journal of Accounting and Economics*, 70(1):101334.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. 2019. Time2vec: Learning a vector representation of time. *CoRR*, abs/1907.05321.

Raehyun Kim, Chan Ho So, Minbyul Jeong, Sanghoon Lee, Jinkyu Kim, and Jaewoo Kang. 2019. HATS: A hierarchical graph attention network for stock movement prediction. *CoRR*, abs/1908.07999.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Manfred Klenner, Don Tuggener, and Simon Clematide. 2017. Stance detection in facebook posts of a german right-wing party. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, LSDSem@EACL 2017, Valencia, Spain, April 3, 2017*, pages 31–40. Association for Computational Linguistics.

Olga Kolchyna, Tharsis T. P. Souza, Philip Treleaven, and Tomaso Aste. 2015. Twitter sentiment analysis: Lexicon method, machine learning method and their combination.

Patty Kostkova, Vino Mano, Heidi J. Larson, and William S. Schulz. 2017. Who is spreading rumours about vaccines? influential user impact modelling in social networks. In *Proceedings of the 2017 International Conference on Digital Health*, DH '17, page 48–52, New York, NY, USA. Association for Computing Machinery.

Srikumar Krishnamoorthy. 2018. Sentiment analysis of financial news articles using performance indicators. *Knowledge and Information Systems*, 56(2):373–394.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Mark Lang and Lorien Stice-Lawrence. 2015. Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics*, 60(2-3):110–135.

Jonathan Law. 2018. *A dictionary of finance and banking*, 6 ed. edition. Oxford quick reference. Oxford University Press, Oxford, England.

Jaejun Lee, Raphael Tang, and Jimmy Lin. 2019. What would elsa do? freezing layers during transformer fine-tuning. *CoRR*, abs/1911.03090.

Sang Il Lee and Seong Joon Yoo. 2020. Multimodal deep learning for finance: integrating and forecasting international stock markets. *The Journal of Supercomputing*, 76(10):8294–8312.

Els Lefever and Véronique Hoste. 2016. A classification-based approach to economic event detection in dutch news text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Xiaodong Li, Haoran Xie, Ran Wang, Yi Cai, Jingjing Cao, Feng Wang, Huaqing Min, and Xiaotie Deng. 2016. Empirical analysis: stock market prediction via extreme learning machine. *Neural Comput. Appl.*, 27(1):67–78.

Yingjie Li and Cornelia Caragea. 2021. Target-aware data augmentation for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1850–1860, Online. Association for Computational Linguistics.

Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu. 2021. Target-adaptive graph for cross-target stance detection.

Anders Edelbo Lillie and Emil Refsgaard Middelboe. 2019. Fake news detection using stance classification: A survey. *CoRR*, abs/1907.00181.

Jiexi Liu and Songcan Chen. 2019. Non-stationary multivariate time series prediction with selective recurrent neural networks. In *Pacific Rim International Conference on Artificial Intelligence*, pages 636–649. Springer.

Tim Loughran and Bill McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Matthew Ma and Feng Zhang. 2016. Investor reaction to merger and acquisition rumors. *Available at SSRN 2813401*.

Burton G Malkiel. 2003. The efficient market hypothesis and its critics. *Journal of economic perspectives*, 17(1):59–82.

Razan Masood and Ahmet Aker. 2018. The fake news challenge: Stance detection using traditional machine learning approaches. In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2018, Volume 3: KMIS, Seville, Spain, September 18-20, 2018*, pages 126–133. SciTePress.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 31–41.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Techn.*, 17(3):26:1–26:23.

Mitra Mohtarami, James R. Glass, and Preslav Nakov. 2019. Contrastive language adaptation for cross-lingual stance detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4441–4451. Association for Computational Linguistics.

Ryan Moriarty, Howard Ly, Ellie Lan, and Suzanne K. McIntosh. 2019. Deal or no deal: Predicting mergers and acquisitions at scale. In *2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, December 9-12, 2019*, pages 5552–5558. IEEE.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines. *CoRR*, abs/2006.04884.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Huihui Ni, Shuting Wang, and Peng Cheng. 2021. A hybrid approach for stock trend prediction based on tweets embedding and historical prices. *World Wide Web*, pages 1–20.

Mahla Nikou, Gholamreza Mansourfar, and Jamshid Bagherzadeh. 2019. Stock price prediction using deep learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management*, 26(4):164–174.

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*, 73:125–144.

Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. 2016. Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*, pages 1345–1350. IEEE.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rohan Pimprikar, S Ramachandran, and K Senthilkumar. 2017. Use of machine learning algorithms and twitter sentiment analysis for stock market prediction. *International Journal of Pure and Applied Mathematics*, 115(6):521–526.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1778–1789. Association for Computational Linguistics.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, pages 1003–1012.

Mohammed Qasem, Ruppa Thulasiram, and Parimala Thulasiram. 2015. Twitter sentiment classification using machine learning techniques for stock markets. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 834–840. IEEE.

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.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics.

Gabriele Ranco, Darko Aleksovski, Guido Caldarelli, Miha Grčar, and Igor Mozetič. 2015. The effects of twitter sentiment on stock price returns. *PloS one*, 10(9):e0138441.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 338–348. Association for Computational Linguistics.

Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Alexander Dür, Linda Andersson, and Allan Hanbury. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based IR models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1712–1721, Vancouver, Canada. Association for Computational Linguistics.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2020. Poor man's BERT: smaller and faster transformer models. *CoRR*, abs/2004.03844.

T. Y. S. S. Santosh, Srijan Bansal, and Avirup Saha. 2019. Can siamese networks help in stance detection? In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, COMAD/CODS 2019, Kolkata, India, January 3-5, 2019*, pages 306–309. ACM.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1668–1678. Association for Computational Linguistics.

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

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

Ramit Sawhney, Mihir Goyal, Prakhar Goel, Puneet Mathur, and Rajiv Ratn Shah. 2021b. 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. 2020b. Multimodal multi-task financial risk forecasting. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 456–465. ACM.

Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Ratn Shah. 2021c. FAST: Financial news and tweet based time aware network for stock trading. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2164–2175, Online. Association for Computational Linguistics.

Carolina Scarton, Diego F. Silva, and Kalina Bontcheva. 2020. Measuring what counts: The case of rumour stance classification. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 925–932. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI-Künstliche Intelligenz*, pages 1–13.

Robert P. Schumaker and Hsinchun Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, 27(2):12:1–12:19.

Isabel Segura-Bedmar. 2018. Labda's early steps toward multimodal stance detection. In *IberEval@ SEPLN*, pages 180–186.

Sreelekshmy Selvin, R. Vinayakumar, E. A. Gopalakrishnan, Vijay Krishna Menon, and K. P. Soman. 2017. Stock price prediction using lstm, RNN and cnn-sliding window model. In *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017, Udupi (Near Mangalore), India, September 13-16, 2017*, pages 1643–1647. IEEE.

Deven Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5248–5264. Association for Computational Linguistics.

Maria Skeppstedt, Andreas Kerren, and Manfred Stede. 2017. Automatic detection of stance towards vaccination in online discussion forums. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media, DDDSM@IJCNLP 2017, Taipei, Taiwan, November 27, 2017*, pages 1–8.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 226–234, Suntec, Singapore. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3645–3650. Association for Computational Linguistics.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In (Bender et al., 2018), pages 2399–2409.

Mariona Taulé, Francisco M Rangel Pardo, M Antònia Martí, and Paolo Rosso. 2018. Overview of the task on multimodal stance detection in tweets on catalan# 1oct referendum. In *IberEval@ SEPLN*, pages 149–166.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3346–3359. 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.

Jannis Vamvas and Rico Sennrich. 2020a. X -stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference and the 16th Conference on Natural Language Processing, SwissText/KONVENS 2020, Zurich, Switzerland, June 23-25, 2020 [online only]*, volume 2624 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jannis Vamvas and Rico Sennrich. 2020b. X-stance: A multilingual multi-target dataset for stance detection. *CoRR*, abs/2003.08385.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Keenan Venuti. 2021. Predicting mergers and acquisitions using graph-based deep learning. *CoRR*, abs/2104.01757.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the Workshop on Language Technologies and Computational Social Science@ACL 2014, Baltimore, MD, USA, June 26, 2014*, pages 18–22. Association for Computational Linguistics.

Ran Wang, Haibo Su, Chunye Wang, Kailin Ji, and Jupeng Ding. 2019. To tune or not to tune? how about the best of both worlds? *CoRR*, abs/1907.05338.

Zhen Wang, Qiansheng Wang, Chengguo Lv, Xue Cao, and Guohong Fu. 2020. Unseen target stance detection with adversarial domain generalization. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–8. IEEE.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science, NLP+CSS@EMNLP 2016, Austin, TX, USA, November 5, 2016*, pages 138–142. Association for Computational Linguistics.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in NLP. *CoRR*, abs/2101.11974.

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 778–783. Association for Computational Linguistics.

Junchi Yan, Shuai Xiao, Changsheng Li, Bo Jin, Xiangfeng Wang, Bin Ke, Xiaokang Yang, and Hongyuan Zha. 2016. Modeling contagious merger and acquisition via point processes with a profile regression prior. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2690–2696. IJCAI/AAAI Press.

Linyi Yang, Eoin Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. 2020. Generating plausible counterfactual explanations for deep transformers in financial text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6150–6160, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959.

Amir Zadeh, Louis-Philippe Morency, Paul Pu Liang, and Soujanya Poria, editors. 2020. *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*. Association for Computational Linguistics, Seattle, USA.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3188–3197. Association for Computational Linguistics.

Ran Zhao, Yuntian Deng, Mark Dredze, Arun Verma, David Rosenberg, and Amanda Stent. 2019. Visual attention model for cross-sectional stock return prediction and end-to-end multimodal market representation learning. In *The Thirty-Second International Flairs Conference*.

Elena Zotova, Rodrigo Agerri, Manuel Núñez, and German Rigau. 2020. Multilingual stance detection in tweets: The catalonia independence corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 1368–1375. European Language Resources Association.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In *COLING 2016, 26th International Conference on Computational Linguistics, December 11-16, 2016, Osaka, Japan*, pages 2438–2448. ACL.

A Data Analysis

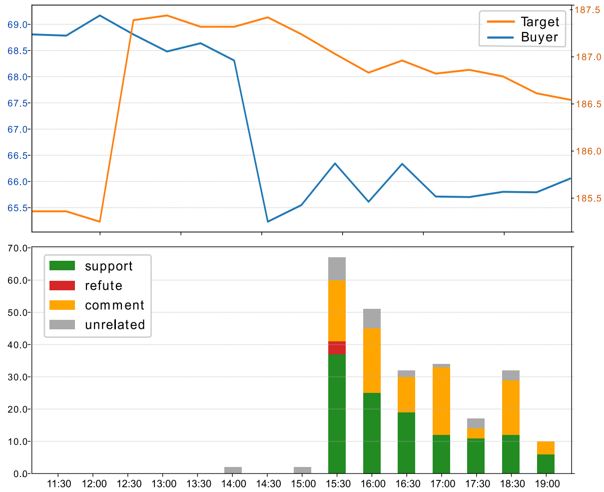

Figure 3: Stock prices of CVS (buyer) and AET (target) on the day of the merger announcement (26.10.2017).

In addition to the example discussed in Section 4, we report a further case study from financial data aligned to WT–WT, this time from one of the succeeded mergers, CVS_AET. As shown in Figure 3, on the day in which the CSV_AET merger was officially announced, the buyer's price decreased, while the target's price increased. This is in line with the theory (Bruner and Perella, 2004) and also makes intuitively sense: the deal was worth $69 billion and CVS was likely to need to pay a premium to acquire AET's shares.

This knowledge is captured by the stock market's movements, and constitutes very valuable information for a stance classifier, as it implicitly increases the likelihood of a *supporting* stance. The lower plot in Figure 3 shows not only a peak in the tweets number, but also in the relative proportion of *supporting* tweets.

B Experimental Specification

B.1 Detailed Data Preprocessing

We perform minimal preprocessing on the textual input: differently than in the BerTweet paper (Nguyen et al., 2020), we perform only URL normalization and lowercasing. We leave the usernames as in their original form: this was done because, in many cases, the usernames are the only clue in the tweet that points to one of the considered companies. To create the string representation for the target, we follow Conforti et al. (2020b)'s representation of company names and acronyms, and add the official (at the time of data collection)

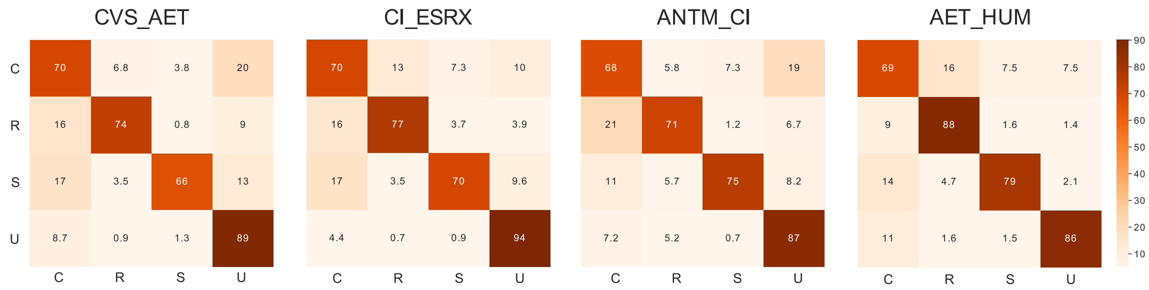

Figure 4: Confusion matrices for the our multi-modal model on the test merger (when training, in turn, on the other three). y axis are the true, and the x the predicted labels, in the order: **C**omment, **R**efute, **S**upport, **U**nrelated.

Twitter account(s) for both the buyer and the target (Table 8).

Company	Acronym	Twitter Username(s)
Aetna	AET	*@Aetna @AetnaHelp*
Anthem	ANTM	*@AnthemInc @Anthem*
Cigna	CI	*@Cigna*
CSV	CVS	*@cvs @cvshealth*
Express Script	ESRX	*@ExpressScripts*
Humana	HUM	*@Humana*

Table 8: Company-related specifications used to obtain the targets.

B.2 Experimental Setup

(Number of) Hyper-Parameters. All models use Adam (Kingma and Ba, 2014) with weight decay $3e-5$, $\beta1 = 0.9$, $\beta2 = 0.999$. Models are trained for a maximum of 7 epochs, with early stopping monitoring the eval loss with a patience of 3. All hyper-parameters used are reported in Table 9 and have been optimized on the development set. Table 10 reports on the total number of (trainable) parameters for each considered model.

batch size	64
maximum tweet length	64
output of BerTweet	768
financial input vector size	12
financial input sequence length	30
GRU hidden size	128
number of attention heads	6

Table 9: Details of used hyper-parameters.

Training Setting. All models are trained using cross-validation, testing on one target and training on the other three. The WT–WT dataset does not provide any official development set. Following (Conforti et al., 2020b), we randomly select a 15% of the training sample as development set.

Model	#parameters	#trainable parameters
BerTweet (frozen)	134,903,044	49,848,580
BerTweet (frozen:9)	"	71,112,196
BerTweet (trained)	"	134,903,044
SDTF (MT, frozen)	168,783,423	83,727,167
SDTF (MT, frozen:9)	"	104,992,575
SDTF (MT, trained)	"	168,783,423

Table 10: Number of (trainable) parameters for all considered models and training settings.

Evaluation Framework. We use sklearn's implementation[17] of accuracy and macro-averaged precision, recall and F_1 scores (Pedregosa et al., 2011).

Computing Infrastructure and Runtime Specifications. Models were trained on Google Colab's GPU. On average, each experiment took ∼1:30 hours to train.

Confusion Matrices. Detailed confusion matrices for all cross-validation settings are reported in Figure 4.

[17] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics