Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification

Linyi Yang[1], **Eoin M. Kenny**[1], **Tin Lok James Ng**[2], **Yi Yang**[3], **Barry Smyth**[1], and **Ruihai Dong**[1]

[1]Insight Centre ,University College Dublin, Ireland
[1]{first.last}@insight-centre.org
[2]University of Wollongong, Australia
[2]jamesng@uow.edu.au
[3]The Hong Kong University of Science and Technology, Hong Kong, China
[3]imyiyang@ust.hk

Abstract

Corporate mergers and acquisitions (M&A) account for billions of dollars of investment globally every year, and offer an interesting and challenging domain for artificial intelligence. However, in these highly sensitive domains, it is crucial to not only have a highly robust and accurate model, but be able to generate useful explanations to garner a user's trust in the automated system. Regrettably, the recent research regarding eXplainable AI (XAI) in financial text classification has received little to no attention, and many current methods for generating textual-based explanations result in highly implausible explanations, which damage a user's trust in the system. To address these issues, this paper proposes a novel methodology for producing plausible counterfactual explanations, whilst exploring the regularization benefits of adversarial training on language models in the domain of FinTech. Exhaustive quantitative experiments demonstrate that not only does this approach improve the model accuracy when compared to the current state-of-the-art and human performance, but it also generates counterfactual explanations which are significantly more plausible based on human trials.

1 Introduction and Related Work

In recent years, large-scale, pre-trained transformer models have led to massive improvements on a wide range of natural language processing (NLP) tasks (Devlin et al., 2018; Liu et al., 2019), including financial technology applications (Duan et al., 2018; Yang et al., 2018; Xing et al., 2019; Yang et al., 2020). However, this impressive ability also coincides with an inherent lack of *robustness* and *transparency*, which undermines human trust in the prediction outcome. In the highly sensitive (and financially lucrative) area of FinTech, explainable financial text classification remains an open, and highly alluring question. To tackle this problem, this paper advances a novel approach which first applies robust transformer models (by leveraging adversarial training) on a real-world, up-to-date, self-collected mergers and acquisitions (M&A) dataset, and then generating plausible, *post-hoc*, counterfactual explanations. In the remainder of this section, we describe relevant work to both of these areas before detailing our contributions.

1.1 Artificial Intelligence in Mergers and Acquisitions

M&As have reshaped the global business landscape for generations, and are having an accelerating impact on the world's economy as new technologies such as the internet, big data, and artificial intelligence disrupt many business sectors (Yan et al., 2016). To appreciate this, a recent economic study provided strong evidence that M&A deal rumours could influence the share price volatility of rumor target firms (Ma and Zhang, 2016). In particular, they showed that, on average, M&A rumors have a positive short term impact and a negative long term impact on the cumulative abnormal returns of the potential acquirers and targets. In the existing AI literature, focus here is typically on predicting likely M&A targets (Yan et al., 2016), and forecasting the likely success of M&A (Danbolt et al., 2016) for developing high-risk/high-reward investment strategies based on M&A speculation (Ji and Jetley, 2009).

While the existing literature typically focuses on predicting likely M&A acquirers and targets, in this work we address a distinct but related task: namely, whether a merger and acquisition *rumor* is likely going to prove to be correct.

1.2 Visualization-based Explanations

To interpret a model's prediction, prior efforts have focused on either incorporating *pre-hoc* analysis into the experimental design (Brunner et al., 2020), or developing *post-hoc* analysis algorithms to select or modify particular instances of the dataset to explain the behavior of models (Keane and Smyth, 2020; Kenny and Keane, 2019). Recent research (Grimsley et al., 2020) shows that transformer models can not be perfectly explained from their intrinsic architecture, and a further work (Brunner et al., 2020) provides strong evidence that self-attention distributions are not directly interpretable. For this reason, model-agnostic, *post-hoc* explanation methods have come to the fore among these works for explaining text classification models, as they are easy to understand and do not require access to the data or the model (Keane and Smyth, 2020).

Towards *post-hoc* explanation in NLP tasks, (Murdoch et al., 2018) proposes a popular way named contextual decomposition (CD) to quantify the importance of each individual word/phrase by computing the change to the model prediction when solely removing a word/phrase. Its hierarchical extensions (Singh et al., 2019; Jin et al., 2020) continue to refine the explanation algorithms that calculate and further visualize the individual phrase's importance. However, despite these visualization-based methods (Murdoch et al., 2018; Singh et al., 2019; Jin et al., 2020) having achieved good results on a popular dataset of sentiment analysis (namely the Stanford Sentiment Treebank-2 [SST-2] dataset where human create the ground truth with their subjective judgement), how to generate explanations in more complex scenarios where human performance is worse than a model have not been well studied. As a result, the prior lines of visualization-based works cannot provide a clear boundary between positive and negative instances to human, whereas counterfactuals could provide "human-like" logic to show a modification to the input that makes a difference to the output classification (Byrne, 2019). Hence, *post-hoc*, example-based explanation methods have received more and more attention in recent years (Keane and Smyth, 2020).

1.3 Counterfactual Text Explanations

Counterfactual explanations are renown for their explanatory ability in AI systems (Wachter et al., 2017); specifically, they offer the ability to explain models (such as transformers) without having to "open the black-box" (Grath et al., 2018), by conveying causal information about what contributed to a given classification. To understand counterfactuals in the context of text classification, consider a sentiment classification task were a black-box model may classify "John loved the film" with a positive sentiment, and explain the prediction *counterfactually* by presenting "John *hated* the film". Glossed, this latter text is the AI explaining the prediction by saying "f the word *love* was replaced with the word *hate*, I would have thought it was a negative sentiment". This allows us to understand the main reasoning process behind the classifier in question, thus explaining the prediction causally. To understand the issue of counterfactual *plausibility*, consider that the previous explanation may also generate a counterfactual which reads "John *not* the film". This text may "flip" the classification to the counterfactual class, but it is *grammatically implausible*, and (arguably) very difficult to contextualize. The reason this is important is because humans avoid creating counterfactuals which are far from a "possible world" (Wachter et al., 2017), and by extension wildly implausible (Byrne, 2019; Kenny and Keane, 2020). In response to this, our work attempts to guarantee more grammatically plausible explanations, and does not rely on attention weights, nor is it constrained to a specific text domain.

Contributions and Paper Outline

- We present a novel dataset to the interesting and challenging problem of artificial intelligence in M&A prediction.

- To the best of our knowledge, the present work is the first general approach to generate grammatically plausible counterfactual explanations for unstructured text classification.
- The primary technical contribution in this work is to generate grammatically *plausible* counterfactuals by replacing the most important words with the antonyms (REP-SCD) based on pre-trained language models. Furthermore, two additional variants (removing/inserting works at the most important place, namely RM-SCD and INS-SCD) are proposed to guarantee counterfactual generations, albeit ones which are less plausible.

The remainder of this paper is organized as follows. Section 2 details our novel dataset and the pre-processing steps involved. Section 3 describes our adversarial training approach, with the sensitivity-based method for counterfactual explanation generation. Exhaustive experiments (both quantitative and human-based) show clear improvements in our method over current state-of-the-art, both in regards to classification accuracy, and explanation quality (see Sections 4 and 5). Finally, the implications of this work on XAI and future research is discussed.

2 The Novel Mergers and Acquisitions Dataset

Description	Number
#Processed deal news total (2007-2019)	4,098
#Train (2007-2014)	3,120
#Validation (2015 - 2016)	478
#Test (2017 - Aug 2019)	500
#Unique companies and institutions	1,406

Table 1: The description of our dataset

For this study we adopted a large-scale, up-to-date M&A dataset collected from Zephyr, a comprehensive database of deal data from the "real world". The dataset [1] contains 14,539 news articles or tweets on M&A events between January 1st 2007, and August 12th 2019. Each instance corresponds to a specific editorial M&A article which describes a possible deal between an acquirer and a target company (also including a few IPO rumours). Additionally, each datapoint also includes the deal outcome (see below), and the deal announcement data, if relevant. In this work, the deal outcome corresponds to the target class, and the raw dataset contains the following outcome types: *complete* – a deal between the acquirer and target companies concluded successfully; *rumour* – no deal materialized between the acquirer and target company; *pending* – a desired deal between the acquirer and target company has been confirmed, and at the time of data collection was deemed to be in-progress, but not yet complete; *cancelled* – a past potential deal between the acquirer and target companies has been confirmed, but it did not complete, and is no longer being pursued.

In order to prepare the raw dataset for use in this study, a number of pre-processing steps were carried out:

1. In this work we chose to focus on a binary classification task and, as such, removed instances with outcome types of *cancelled* and *pending*, leaving only those instances that correspond to *completed* deals (the positive class) and *rumours* (the negative class).
2. We eliminated instances where *both* acquiring and target companies were non-US, due to a tendency towards low-quality data; in other words, all of the instances in our dataset include a US Listed Company as either the acquirer or the target or both.
3. Articles published within one day or after the deal announcement date were also removed, this is because our interest is in developing a prediction model that is capable of generating accurate predictions at least one day in advance of any deal outcome.

[1] https://www.bvdinfo.com/en-gb/our-products/data/specialist/zephyr

4. Finally, the remaining instances are randomly over-sampled to ensure an even split between positive (completed) and negative (rumours) instances for each year.

The result is a dataset of 4,098 instances (news articles and meta-data) which we split into training, validation, and testing sets on a year-by-year basis (see Table 1).

3 Methodology

The pipeline of our method is shown in Fig. 1. First, as a prerequisite, a transformer variant is fine-tuned on the M&A prediction task, alongside adversarial training (which as we shall see is shown to be promising in this domain). Second, important words in the test instances are identified using a sampled contextual decomposition technique after the prediction. Third, a counterfactual explanation is generated by replacing these words with *grammatically plausible* substitutes. As we shall see, although this method does not always guarantee a plausible counterfactual will be found, we propose two alternative methods which will, albeit with the possible trade-off of plausibility. These steps are detailed next.

3.1 Step 1: Robust Transformer Classification Models

As eluded to earlier, M&A prediction is a highly sensitive domain, and despite adversarial training showing promise previously (Goodfellow et al., 2014; Tsipras et al., 2018), it has never been tested in this domain. Hence, to try ensure a robust model which can simultaneously generate intelligible explanations, we explore its usage here compared to other popular approaches. Given a news article, we adopt the classical transformer architecture proposed by (Vaswani et al., 2017). The original multi-head self-attention is subsequently applied to the k-th document $\mathcal{D}^{(k)}$, which is calculated as follows:

$$\text{MultiHead} = \textbf{Concat}\left(\text{head}_1, \ldots, \text{head}_\text{h}\right) W^O \tag{1}$$

$$\text{head}_\text{j} = \text{Attention}\left(Q, K, V\right) \tag{2}$$

$$Q = \mathcal{D}^{(k)} W_j^Q, K = \mathcal{D}^{(k)} W_j^K, V = \mathcal{D}^{(k)} W_j^V \tag{3}$$

where $W_j^Q, W_j^K, W_j^V \in \mathbb{R}^{d \times d}$ are weight metrics, and the attention is computed as:

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)\mathbf{V} \tag{4}$$

for input query, key and value matrices $Q, K, V \in \mathbb{R}^{n \times d}$. The h outputs from the attention calculations are concatenated and transformed using an output weight matrix $W^o \in \mathbb{R}^{dh \times d}$.

Additionally, the adversarial noise, treated as a form of regularization, is generated by the Fast Gradient Method (FGM) (Miyato et al., 2017) and Projected Gradient Descent (PGD) (Madry et al., 2018). The idea of using adversarial perturbation is derived from the usage of adversarial attacks (Carlini and Wagner, 2017) to evaluate the robustness of neural networks, while the recent advances of using the adversarial training in NLP models (Liu et al., 2020) inspires us to use it as a way of regularization. For each embedded word e in k-th news article $\mathcal{D}^{(k)}$, the FGM computes its perturbation as follows:

$$r_{fgm} = \epsilon \cdot g / \|g\|_2 \text{ where } g = \nabla_e L(\theta, (\mathcal{D}^{(k)}, y)) \tag{5}$$

where r_{fgm} is the perturbation of e, θ denotes the current values of the parameters of the classifier, and L denotes the loss function (cross entropy) associated with the classifier. The perturbation can be easily computed using back-propagation. The projected gradient descent, which can be considered as a multi-step variant of the FGM, computes the perturbation of e iteratively:

$$\begin{aligned} e_{t+1} &= \Pi_{e+S}\left(e_t + \alpha g\left(\mathcal{D}_t^{(k)}\right) / \left\|g\left(\mathcal{D}_t^{(k)}\right)\right\|_2\right) \\ g\left(\mathcal{D}_t^{(k)}\right) &= \nabla_e L\left(\theta, (\mathcal{D}_t^{(k)}, y)\right) \end{aligned} \tag{6}$$

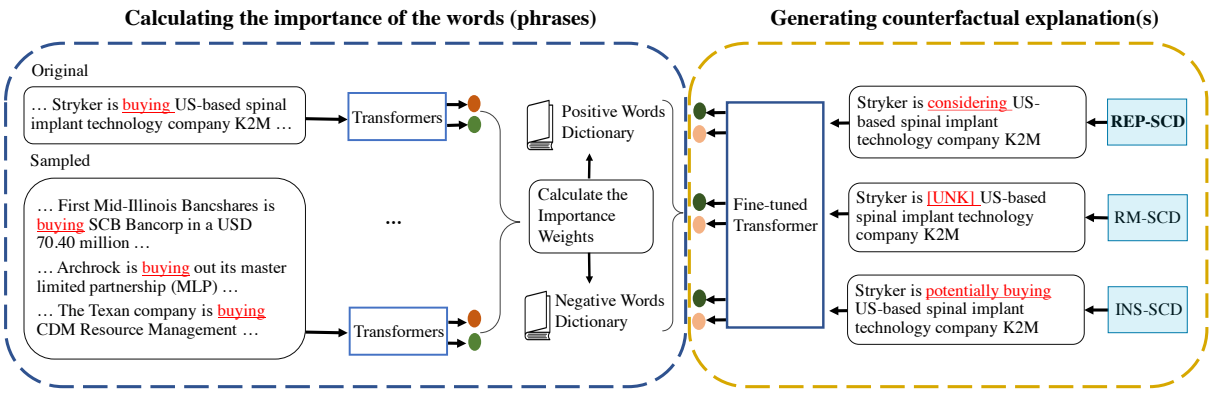

Figure 1: The pipeline of our methods, namely REP-SCD, RM-SCD, and INS-SCD. We show real examples of generating diverse counterfactual instances that flip the prediction result from *completed* to *rumour*. The original input has been changed by iteratively modifying words in order of their importance until the prediction matches the counterfactual class. The outputs (logits) of the predictions are represented in green, and orange points, respectively.

where $S = \left\{r \in \mathbb{R}^d : \|r\|_2 \leq \epsilon\right\}$ is the constraint space of the perturbation, Π_{e+S} denotes the projection of a vector onto the feasible set $e + S$, and α is the step size. We use Adam optimizer with learning rate decay to train our model until convergence.

3.2 Step 2: Context-Independent Word Importance

To calculate the context independent importance up to one word, we adopt the *sensitivity of contextual decomposition* technique from (Madry et al., 2018) which removed part of inputs from the sequence text to evaluate a model's sensitivity to them, thereby allowing for the identification of important features. In its hierarchical extensions – Sampling and Contextual Decomposition (SCD), (Jin et al., 2020) mask out the phrase p from the input while the max sequence length N is set to 40. However, the average input length in our data is much larger than 40. We, therefore, propose a phrase-level removing method only if the phrase starts with the negative pronouns or limitations. Otherwise, only a single word will be removed. For example, in the sentence "the deal is not closing currently", the attribution of "closing" should be positive while the attribution of "not closing" should be negative. In this situation, we remove the whole phrase "not closing" together to calculate the influence in terms of the logits change in the output layer of the transformer and then assign the negative score to the word "closing".

Given a phrase p starting with the negative limitations in the k-th document $\mathcal{D}^{(k)}$, we sample the documents which contain the same phrase p to alleviate the influence by chance when there are multiple shreds of evidence saturating the prediction. For example, in the source "JPMorgan is closing in on a deal, sources close to the situation are optimistic for deal completion", if we only remove the word "closing", the prediction would not be changed so much. In this sampling way, the proposed context-independent importance of word and phrase is more robust to saturation. The formula for calculating the importance can be written as:

$$\phi(\mathbf{p}, \widehat{\mathcal{D}^{(k)}}) = \mathbb{E}_{\widehat{\mathcal{D}^{(\beta)}}}\left[l\left(\widehat{\mathcal{D}^{(\beta)}}; \widehat{\mathcal{D}}\right) - l\left(\widehat{\mathcal{D}^{(\beta)}} \backslash \mathbf{p}; \widehat{\mathcal{D}}\right)\right] \tag{7}$$

where $\mathcal{D}^{(\beta)}$ denotes the resulting document after masking out a single token or a phrase starting with the negative pronoun in the length of N surrounding the phrase $\mathbf{p}$. we use $l\left(\widehat{\mathcal{D}^{(\beta)}} \backslash \mathbf{p}; \widehat{\mathcal{D}}\right)$ to represent the model prediction logits after replacing the masked-out context. $\backslash \mathbf{p}$ indicates the operation of masking out the phrase p in a input file sampling from the testing set $\mathcal{D}$.

As an aside, the resulting top 15 most influenced words are shown in Table 2. In total, there are 123 positive words and 155 negative words in the dictionaries. We can see the average influence score of positive words (0.637) is higher than the negative words (0.385). It may reveal that positive words usually contain more powerful clues in predicting the M&A deal. That would be interesting to see which kind of words in the sources illustrate the deal is more likely to be completed in the future and which kind of words would be likely to kill the deal.

Algorithm 1 Plausible Counterfactual Instances Generation

Input: Testing document example $\mathcal{D}^{(k)}$= $\{w_1, w_2, ..., w_n\}$, the corresponding ground truth label Y, pre-trained Mask Language Model MLM, negative pronouns list NP, fine-tuned transformer classifier C.
Output: Positive Word Dictionaries POS, Negative Word Dictionaries NEG, Plausible counterfactual example(s) $D_{cf}^{(k)}$= $\{D_{REP-SCD}^{(k)}, D_{RM-SCD}^{(k)}, ..., D_{INS-SCD}^{(k)}\}$

Initialization: $D_{cf}^{(k)} \leftarrow D^{(k)}$
for each word w_i in in $\mathcal{D}^{(k)}$ **do**
  if the prev word w_{i-1} is in NP **then**
    Creat the whole phrase np_i by contextual decomposition
    Computer the importance score $P_{w_i} = -P_{np_{(i)}}$ via Eq.(7)
  else
    Computer the importance score P_{w_i} via Eq.(7)
  end if
end for
Create dictionaries with words: $W_{POS}; W_{Neg}$, alongside the word positions pos_{w_i} sorted by the descending order of their importance scores P_{w_i}.
for each word position pos_i in pos_{w_i} **do**
  $W_{Plausible} \leftarrow MLM(D_{mask\_w_{pos_i}}^{(k)})$, $W_{Plausible}^{'} \leftarrow MLM(D_{mask\_w_{pos_i \pm 1}}^{(k)})$
  if $Y^{(k)}$ == POS **then**
    $W_{Candidate}$, $W_{Candidate}^{'} \leftarrow$ Intersection (W_{NEG} and $W_{Plausible}$), (W_{NEG} and $W_{Plausible}^{'}$)
  else
    $W_{Candidate}$, $W_{Candidate}^{'} \leftarrow$ Intersection (W_{POS} and $W_{Plausible}$), (W_{POS} and $W_{Plausible}^{'}$)
  end if
  $D_{rm}^{(k)} \leftarrow D^{(k)}\_w_{pos_i}$
end for
for each word w_i,$w_i^{'}$ in zip ($W_{Candidate}$,$W_{Candidate}^{'}$) **do**
  $D_{ins}^{(k)} \leftarrow$ Insert $w_i^{'}$ to $D_{mask\_w_{pos_i \pm 1}}^{(k)}$
  $D_{rep}^{(k)} \leftarrow$ Replace w_i with $D_{mask\_w_{pos_i}}^{(k)}$
  if $C(D_{rm}^{(k)}, D_{ins}^{(k)}, D_{rep}^{(k)}) \neq$ Y **then**
    Add $D_{rm}^{(k)}, D_{ins}^{(k)}, D_{rep}^{(k)}$ to the set $D_{cf}^{(k)}$
  end if
end for
return $D_{cf}^{(k)}$

3.3 Step 3: Counterfactual Instance Generation

As shown in Algorithm 1, we summarize three different counterfactual generation methods, namely, the primary technique which generates grammatically plausible counterfactuals (REP-SCD), and two further variants to guarantee counterfactual generation (RM-SCD and INS-SCD). We combine these three methods to alleviate a major issue in counterfactual explanation, that is, there is no guarantee that for a given example a counterfactual instance is found. Our main technique identifies the most important word(s) in a test instance using SCD and replaces them with the intersection of grammatically plausible substitutes [using masked language model (MLM)] and words in the reverse emotional dictionary. The raw document content $\mathcal{D}^{(k)}$ itself is taken as input, and MLM outputs $p(\cdot|\mathcal{D}^{(k)})$ for each masked position. After all masked positions are infilled, we get the reconstructed document:

$$\widehat{\mathcal{D}^{(k)}} = \mathrm{MLM}(\mathcal{D}^{(k)}). \tag{8}$$

We iterative repeat this operation at the most important word positions ranked by SCD until the reconstructed document ultimately moves the model's classification towards the opposing class. Notably, there

Positive Words	Sensitivity	Negative Words	Sensitivity
announced	5.841	talks	4.674
line	5.715	could	2.484
announcement	4.469	flag	2.236
agreement	3.378	diligence	1.363
acquiring	3.342	considering	1.196
completion	2.727	time	1.186
agreed	2.429	may	1.085
closing	2.125	looking	0.983
consideration	1.994	this	0.972
prevailed	1.639	when	0.914
acquire	1.520	potentially	0.870
paid	1.461	if	0.847
disclosed	1.403	intention	0.836
selling	1.385	year	0.812
could	1.360	takeover	0.790

Table 2: Top 15 most influenced words towards the M&A prediction. The influence score for each word is calculated and added up by Sampling and Contextual Decomposition (SCD) on the testing set.

may be more than one counterfactual explanation corresponding with the original text instance.

4 Experiment 1: Financial Text Classification with Robust Transformers

In this section we describe the results of a comprehensive evaluation of classification accuracy, comparing a variety of different classification baselines (including a human baseline) to our adversarial transformer approach.

4.1 Methods Used

The baselines used can be grouped into several distinct categories: human evaluations – traditional machine learning approaches (SVM) – classical deep learning approaches (CNN (Kim, 2014), BiGRU (Bahdanau et al., 2014) , and HAN (Yang et al., 2016)) – and various transformer approaches with/without pruning strategies. These transformer-based models are generally considered to provide the current state-of-the-art in text classification. We reproduce these baselines based on the Transformers.[2]

Acquiring a human baseline As a baseline, we asked 26 participants which were experts in economics and finance to predict M&A events by completing 50 M&A evaluation questionnaires. The participants consisted of Ph.D. students, and academics from the fields of economics/finance. All participants were either native English speakers or had a high degree of English competence. Each questionnaire provided information on ten M&A cases/instances, sampled randomly without replacement from the test set. In addition, the news articles available in the dataset that were published before the deal announcement were also provided. The questionnaire asked the participant to predict the outcome of the deal (complete or rumour), and to state their confidence in this prediction.

4.2 Classification Results

In line with best practice, model hyper-parameters are tuned using the validation set. In particular, the maximum sequence length is set as 256, and the size of transformers are all set as large. All experiments are using the conventional Matthews Correlation Coefficient *(MCC)*, accuracy and *F1* metrics. The classification results are summarized in Table 3 with *Random Guess* used to provide a lower-baseline based on chance. While the human evaluators performed better than chance their ability to predict deal outcomes is limited when compared to the more sophisticated machine models that follow. These results are particularly compelling as the human evaluators had considerable domain expertize.

[2]https://github.com/huggingface/pytorch-transformers

Evaluation	MCC	Accuracy	F1	Evaluation	MCC	Accuracy	F1
Baselines				**Transformers**			
Random Guess	0.013	0.510	0.462	ALBERT	0.768	0.882	0.879
Human Evaluation	0.307	0.640	0.672	+Ad.	0.780	0.890	0.888
Traditional ML				DistilBERT	0.750	0.874	0.877
SVM(TF-IDF)	0.701	0.816	0.816	+Ad.	0.784	0.890	0.891
Classical DL				BERT-WWM	0.751	0.874	0.879
CNN-Text	0.729	0.848	0.847	+Ad.	***0.788***	***0.894***	0.894
BiGRU	0.734	0.836	0.849	RoBERTa	0.780	0.892	0.888
HAN	0.742	0.848	0.853	+Ad.	***0.788***	***0.894***	***0.895***

Table 3: Evaluations performed by human, machine learning, deep learning, and transformer-based models, alongside the ablation study for adversarial training (indicate as +Ad.). The scores in bold and italics indicate the best performance across all approaches.

Each of the machine learning approaches offer substantial improvements over the human evaluators and a clear separation can be seen between traditional machine learning (with MCC scores in low 0.7 range/F1 scores in the low 0.8 range), classical deep learners (with MCC scores in the range 0.73-0.74/F1 scores in the range 0.84-0.85), and recent transformer-based models (MCC$>$0.75/F1$>$0.87).

We further evaluate the relative influence of the adversarial perturbation to test the robustness of the models. We find that all variants of the transformer (Lan et al., 2019; Sanh et al., 2019) benefit from the adversarial perturbation during the training process in terms of the prediction results in the practice. For exploring the reason why the optimal transformer classifier can outperform the human test a lot – 39%, we take the best performed model – RoBERTa (Liu et al., 2019) with adversarial training as our optimal classifier in the following experiments for generating the plausible counterfactual explanations.

5 Experiment 2: Generating Plausible Counterfactual Explanations

Interpretability is an increasingly important property for many deep learning techniques, including computer vision and natural language processing (Kenny and Keane, 2019), especially in critical tasks such as financial text classification; high-value investment decisions demand a reasonable level of interpretability if investors are to trust the predictions that come for a system such as the one described in this work. In this section, we describe the qualitative analysis for each of our methods. Subsequently, we show the evaluation of user studies compared to the existing example-based explanation methods.

5.1 Qualitative Analysis for the Resulting Counterfactual Instances

In qualitative analysis, we identified five typical patterns among the generated counterfactual instances as shown in Table 4 where we highlight the changing parts. Based on the 500 testing examples, we guarantee that there is at least one counterfactual instance corresponding with the original input. We gain insight into which aspects are causally relevant by comparing the original context to the revised context which can flip the classifier's prediction.

5.2 Human Evaluation for the Explanation

We implement interpretation experiments on the optimal fine-tuned transformer classifier. While an explainable model trained with supervised learning is a common method to interpret the results of text classification (Wallace et al., 2019), the self-supervised learning explainable frameworks have been scarcely found. Meanwhile, the work in (Kaushik et al., 2020) consider similar types of edits to generate counterfactually-revised data, however, all of the instances are generated by human which greatly limits the expansibility of the method. To comprehensively evaluate the performance of our method, we consider a state-of-the-art example-based explanation framework for comparison, namely HotFlip (Ebrahimi et al., 2017), which uses gradients to identify important words and then flip it with the adversarial word which can cause the maximum change in gradients.

Types of Algorithms	Examples
REP-SCD: Replacing with the certainty word	Ori: Professional vacation services provider ILG is **considering** a merger with Diamond Resorts International... Rev: Professional vacation services provider ILG is **announcing** a merger with Diamond Resorts International...
REP-SCD: Changing the deal value	Ori: Vivendi is in early discussions to sell a 10.0 per cent stake in Universal Music Group (UMG) to Tencent for roughly EUR 3.00 **billion**... Rev: Vivendi is in early discussions to sell a 10.0 per cent stake in Universal Music Group (UMG) to Tencent for roughly EUR 3.00 **million**
INS-SCD: Recasting $fact$ as $hoped\ for$	Ori: Stryker **is buying** US-based spinal implant technology company K2M Group Holdings for USD 1.40 billion in cash Rev: Stryker **is potentially buying** US-based spinal implant technology company K2M Group Holdings for USD 1.40 billion in cash
INS-SCD: Inserting the negative word	Ori: WPP has **confirmed** the recent speculation that it has entered into exclusive negotiations with private equity firm Bain Capital... Rev: WPP has **not confirmed** the recent speculation that it has entered into exclusive negotiations with private equity firm Bain Capital...
RM-SCD: Removing the negative limitation(s)	Ori: This suitor is the Namdar and Washington Prime consortium, the insiders noted, adding that there can be **no certainty** a deal will complete... Rev: This suitor is the Namdar and Washington Prime consortium, the insiders noted, adding that there can be **certainty** a deal will complete...

Table 4: Most prominent categories of counterfactual explanations generated by our algorithms, namely RM-SCD, REP-SCD, and INS-SCD for M&A Predictions. Ori and Rev are short for original and revised instances, respectively.

For user evaluation, here we ask domain experts in finance to rate our explanations on two aspects, (1) how *plausible* (mainly in terms of grammar and comprehension) it is, and (2) how *reasonable* it is (i.e., does the explanation make sense). We compare our method to Hotflip - the current state-of-the-art framework for counterfactual explanation - at the time of writing. Each score is measured on a scale of 1-5, where 5 is the best, and 1 is the worst. We randomly sample 100 examples from the testing set for 5 participants to answer (20 examples per person). By combining the REP-SCD, RM-SCD, INS-SCD together, our method achieves significantly higher ranking score compared to HotFlip, more specifically, 2.35 score improvements (4.35/2.00) were made regarding plausibility while 0.85 score improvements (4.00/3.15) were made on reasonableness, showing a p-value less than 0.001 and 0.05, respectively. Hence, there is compelling evidence that our method can generate counterfactual explanations which are more *plausible* and *reasonable*.

6 Conclusion and Future Work

In this work, we pursued a new research problem of M&A prediction. Our transformer-based classifier leveraged the regularization benefits of adversarial training to enhance model robustness. More importantly, we built upon previous techniques to quantify the importance of words and help guarantee the generation of plausible counterfactual explanations with a masked language model in financial text classification. The results demonstrate superior accuracy and explanatory performance compared to state-of-the-art techniques. An obvious extension would be to include canceled deals into the classifier,

or to predict novel M&A events based on market descriptions of companies (e.g., scale, finances, and target markets). Moreover, additional financial events (e.g., misstatement detection and earnings call analysis) is yet another related task to be considered for further research.

Acknowledgment

We would like to thank Tianhao Fu, Yimeng Li, Yang Xu and Prof. Mark Keane for their helpful advice and discussion during this work. Also, we would like to thank the anonymous reviewers for their insightful comments and suggestions to help improve the paper. This research was supported by Science Foundation Ireland (SFI) under Grant Number $SFI/12/RC/2289\_2$.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ruth MJ Byrne. 2019. Counterfactuals in explainable artificial intelligence (xai): evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6276–6282.

Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE.

Jo Danbolt, Antonios Siganos, and Abongeh Tunyi. 2016. Abnormal returns from takeover prediction modelling: Challenges and suggested investment strategies. *Journal of Business Finance & Accounting*, 43(1-2):66–97.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Junwen Duan, Yue Zhang, Xiao Ding, Ching Yun Chang, and Ting Liu. 2018. Learning target-specific representations of financial news documents for cumulative abnormal return prediction. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING-18)*, pages 2823–2833.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. 2018. Interpretable credit application predictions with counterfactual explanations. *arXiv preprint arXiv:1811.05245*.

Christopher Grimsley, Elijah Mayfield, and Julia RS Bursten. 2020. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1780–1790.

Xinyu Ji and Gaurav Jetley. 2009. The shrinking merger arbitrage spread: Reasons and implications. *Financial Analysts Journal*, 66, 03.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2020. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mark T Keane and Barry Smyth. 2020. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). In *International Conference on Case-Based Reasoning (ICCBR)*.

Eoin M Kenny and Mark T Keane. 2019. Twin-systems to explain artificial neural networks using case-based reasoning: comparative tests of feature-weighting methods in ann-cbr twins for xai. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2708–2715. AAAI Press.

Eoin M Kenny and Mark T Keane. 2020. On generating plausible counterfactual and semi-factual explanations for deep learning. *arXiv preprint arXiv:2009.06399*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

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

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from LSTMs. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Chandan Singh, W. James Murdoch, and Bin Yu. 2019. Hierarchical interpretations for neural network predictions. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2018. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. Allennlp interpret: A framework for explaining predictions of nlp models. *arXiv preprint arXiv:1909.09251*.

Frank Z Xing, Erik Cambria, and Yue Zhang. 2019. Sentiment-aware volatility forecasting. *Knowledge-Based Systems*, 176:68–76.

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 *International Joint Conferences on Artifical Intelligence, IJCAI-16*, pages 2690–2696.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-16*, pages 1480–1489.

Linyi Yang, Zheng Zhang, Su Xiong, Lirui Wei, James Ng, Lina Xu, and Ruihai Dong. 2018. Explainable text-driven neural network for stock prediction. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pages 441–445. IEEE.

Linyi Yang, Tin Lok James Ng, Barry Smyth, and Ruihai Dong. 2020. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020*, WWW '20, page 441–451.