Classifying US Firms into Trade Peace and Trade War Stocks: an NLP Analysis of Earnings Call Transcripts Approach

Matéo Molinaro - Enzo Montariol

31 July 2025

Contents

1 Abstract

Global trade developments can affect firms' performance, particularly for companies with significant international exposure. Understanding how sensitive individual firms are to trade-related shocks and classifying them as "trade peace" (benefiting from trade policy shocks) or "trade war" stocks is crucial for investors, policymakers, and corporate decision-makers. This study proposes an approach to classify firms' trade sensitivity by applying Natural Language Processing (NLP) techniques to 54,851 earnings call transcripts of Russell 1000 constituents from 2007 to 2025. We employ bag-of-words (BoW), sentiment-adjusted BoW (BoWws), TF-IDF combined with logistic regression, and FinBERT, a transformer-based financial language model, to extract textual signals related to trade exposure. Firms are classified as trade peace or trade war stocks based on model predictions, and we evaluate the economic relevance of these classifications through an event study of cumulative abnormal returns (CARs) around major trade policy events. Our findings suggest that while none of the models clearly discriminate between trade peace and trade war stocks in terms of opposite-signed CAR differences, the TF-IDF–logistic regression model achieves the largest absolute difference in mean cross-sectional CARs, equal to 1.29%. These results highlight the potential of earnings call transcripts as a source for measuring firms' trade sensitivity and provide insights into the strengths and limitations of various NLP-based approaches for classifying firms' exposure to global trade shocks.

2 Introduction

2.1 Companion GitHub

The data, as well as the code and the outputs are available in this GitHub repository

2.2 Problem statement and importance

Global trade can impact firms' performance, particularly for companies with strong exposure to international markets. Understanding how sensitive individual firms are to trade-related developments and being able to classify them between trade peace (firms benefiting from trade policy shocks) and trade war stocks is therefore crucial for investors, policymakers, and corporate decision-makers. In this project, we propose an approach to classify trade peace and trade war stocks by leveraging Natural Language Processing (NLP) techniques on 54851 earnings call transcripts. Specifically, we extract textual information using bag-of-words representations, sentiment-adjusted bag-of-words, and TF-IDF vectorization combined with logistic regression, as well as FinBERT, a transformer-based financial language model.

To evaluate the relevance of our trade-sensitivity classifications, we conduct an event study around predefined periods associated with major trade-related events. Firms are classified as "trade peace" or "trade war" stocks based on the predictions of our models, and we analyze abnormal returns around these events. If our measure successfully captures firms' exposure to global trade, we expect to observe statistically significant differences in abnormal returns between the two groups, with opposite signs for "trade peace" and "trade war" stocks. Such findings would provide empirical validation for the proposed methodology and highlight the value of earnings call transcripts as a rich source of information for assessing firms' trade sensitivity.

3 Data

3.1 Earning calls

To retrieve the earnings call transcripts, we used the Ninjas Earnings Call Transcript API. As the downloading is quite long, we store the transcripts from 2007-05-10 to 2025-07-24 here to just load them when running the code. We split them into chunks to be able to put them in the GitHub repository because large files cannot be stored online.

3.2 Market prices

We focused our analysis on Russell 1000 firms. The historical constituents of the Russell 1000 is available here - RIY Index constituents.feather. The associated stock-level returns are stored in RIY Index returns.feather and are available from 2015-01-05 to 2025-06-11. We'll use the Russell Index (total_returns_russell.feather) as the market as well as the [XX] as the risk-free (rf_returns.csv) to compute abnormal returns.

3.3 Sentiment dictionary

We used the Loughran-McDonald (Loughran-McDonald_MasterDictionary_1993-2024.csv) sentiment dictionary to define positive and negative words.

3.4 Trade policy shocks dates

We use the dates proposed by Scientific Beta supplemented by the recommendations of ChatGPT. Another way of defining the event dates more quantitatively can be to compute rolling percentiles of the TPU Index and identify periods exceeding the rolling percentiles.

Table 1: U.S. Tariff Events (2008–2025)

Date	**Event Description**
12 Oct 2014	U.S. imposes tariffs on Mexican sugar in trade dispute.
24 Apr 2017	U.S. imposes ∼20% countervailing duties on Canadian softwood lumber.
08 Aug 2017	U.S. launches Section 301 investigation into Chinese intellectual property theft.
01 Jan 2018	U.S. imposes 30% tariff on solar panels and 20–50% on washing machines, mainly from China.
01 Mar 2018	U.S. announces 25% tariffs on steel and 10% on aluminum imports (Section 232).
22 Mar 2018	US orders identification of Chinese products for tariffs.
02 Apr 2018	China retaliates with 15–25% tariffs on $3B worth of U.S. goods.
15 Jun 2018	China retaliates on USD50bn of US imports.
17 Sep 2018	Third wave: 10% tariffs (rising to 25%) on approximately $200B of Chinese goods (Phase 3).
10 May 2019	U.S. raises Phase 3 tariffs from 10% to 25% on approximately $200B of Chinese goods.
23 Aug 2019	China raises tariffs on soy and autos
01 Sep 2019	U.S. starts 15% tariffs on approximately $112B more Chinese goods (Phase 4A).
15 Dec 2019	Planned tariffs on remaining $160B delayed or suspended (Phase 4B).
15 Jan 2020	U.S.–China "Phase 1" trade deal signed: partial tariff rollback and new trade commitments.
14 May 2024	Biden administration raises tariffs on EVs, solar cells, steel, aluminum from China.
01 Feb 2025	U.S. imposes 25% tariffs on Canadian and Mexican imports (except 10% on energy).
04 Feb 2025	"Liberation Day Tariffs": U.S. adds 25% tariffs on aircraft, tools, and electronic goods.
04 Mar 2025	China retaliates with approx. 15% tariffs on U.S. agriculture and manufactured products.
12 Mar 2025	U.S. reimposes steel/aluminum tariffs (25%), plus new auto tariffs by early April.
23 Jul 2025	U.S.–Japan agreement: U.S. caps auto tariffs at 15% (down from 25%).

The last event date will not be used in our analysis due to the fact that returns are note available for this period.

4 Classifying stocks as trade peace or trade war on trade policy shock events

4.1 A Bag-Of-Words (BoW) approach

The baseline model we choose is a BoW model for its simplicity and for its use in a related study of Scientific Beta in Firm-level exposure to trade policy shocks: A multidimensional measurement approach, which inspired us heavily for our study, notably for the trade related dictionary, the event dates, as well as the abnormal returns study around events.

$$\textit{Trade score Bow} = \frac{\sum_{w}^{W_T} \left(1_{w \in \text{keywords}}\right)}{W_T} \tag{1}$$

The Trade Bow measure is defined as the fraction of words in the earnings call transcripts that are in the trade dictionary. W_T is the total number of words in a firm's earnings call transcript and $1_{w \in \text{keywords}}$ takes the value of 1 when word w is in the trade dictionary and 0 otherwise.

To construct our classification based on the BoW model, we proceed as follow: first, we preprocess our data by tokenizing, removing stopwords, and stemming (SnowballStemmer) each transcript (using NLTK Python package). Then, at each filing date (date when a transcript becomes available) and for each stock (identified with tickers) we compute the frequency of trade related keywords in the earnings calls transcript of the company. The trade dictionary is based on Baker, Bloom, and Davis (2016) and Caldara et al. (2020) and contains the following keywords:

Trade dictionary	
tariff	World Trade Organization
import duty	trade treaty
import barrier	trade agreement
import ban	trade policy
import tax	trade act
import subsidies	trade relationship
export ban	free trade
export tax	Doha round
export subsidies	Uruguay round
government subsidies	dumping
GATT	border tax
WTO	

Table 2: Trade dictionary terms.

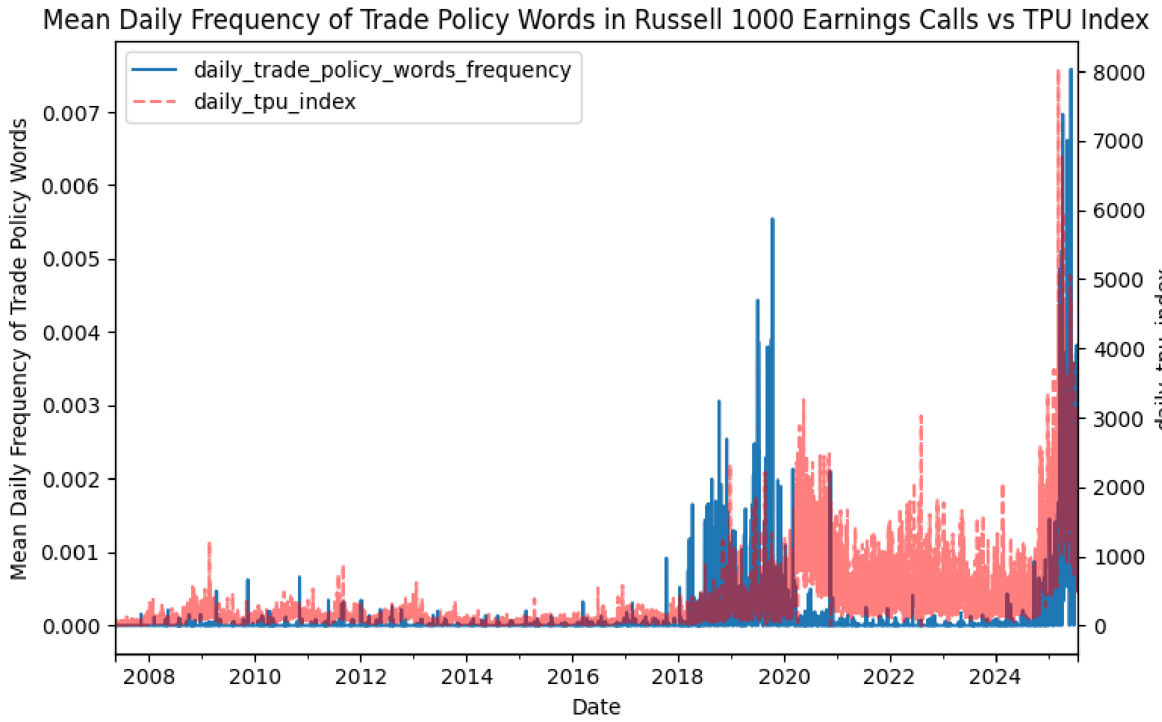

Figure 1: Mean Daily Frequency Trade Policy Words in Russell 1000 Earnings Call Transcripts vs TPU Index

As we can see with the above plot, most of the time, the frequency is 0.0 as trade policy shocks are rare and hence discussions in earnings call are. For this reason, we cannot form classical deciles portfolios as most of the times frequency is null, portfolios cannot be created at these dates and concentration of values around 0.0 may cause problem when computing deciles (not enough unique values). To overcome this issue, one could use 10-Ks as there is a good coverage (Scientific Beta did that, we retrieved the 10-Ks but it will be a future work) but we decided to count the number of stocks mentioning at least one of the keywords in the dictionary and group them by quarter. Thus, we define the trade peace stocks of the BoW model as the firms mentioning at least one keyword in the trade dictionary the quarter preceding the event dates. Inversely, trade war stocks are defined as those not mentioning any keywords. Thus, at each event date, we have our two groups of trade peace and trade war stocks.

We report below, the number of firms mentioning at least a trade keyword by quarter:

Table 3: Number of firms mentioning at least one trade keyword by quarter

Quarter	Number of Firms
2017-01-01	76
2017-04-01	35
2017-10-01	23
2018-01-01	43
2018-04-01	146
2019-01-01	217
2019-04-01	211
2019-07-01	252
2019-10-01	216
2024-01-01	35
2024-10-01	160
2025-04-01	812

4.2 A BoW with sentiment (BoWws) refinement

A dictionary-based approach (BoW) similar to the one we used to construct the trade peace and trade war stock classification of the BoW model (TP&TW BoW) is suitable to capture how much attention firms spend on trade policy in their earnings calls but does not capture the sentiment in these discussions. Therefore, we follow Hassan et al. (2019) in their definition of a measure that also aims to capture the sentiment surrounding a given topic.

$$\textit{Trade sentiment} = \frac{\sum_{w}^{W_T} \left(1_{w \in \text{keywords}} \cdot \sum_{c=w-10}^{w+10} S(c) \right)}{W_T} \tag{2}$$

W_T is the total number of words in an earnings call, $1_{w \in \text{keywords}}$ takes the value 1 if the term w is in the trade dictionary and 0 otherwise, and $S(c)$ takes the value $+1$ when c is in the positive word list from Loughran and McDonald (2011), -1 when c is in their negative word list, and 0 otherwise.

We classify the stocks as TP stocks if the trade sentiment is positive and as TW stocks otherwise. We report below the number of TP&TW BoWws stocks on the quarter preceding the event dates.

Table 4: Number TP and TW stocks of trade keywords at quarter preceding event dates

Quarter	TP	TW
2017-01-01	13	15
2017-04-01	8	7
2017-10-01	1	5
2018-01-01	6	12
2018-04-01	29	47
2019-01-01	58	81
2019-04-01	65	80
2019-07-01	70	89
2019-10-01	71	78
2024-01-01	5	15
2024-10-01	39	57
2025-04-01	247	372

4.3 A TF-IDF - logistic regression model

To capture the relationship between firm-level earnings call content and their potential sensitivity to global trade and classify them between TP&TW stocks, we developed a custom TF-IDF vectorizer combined with a logistic regression classifier.

The approach starts by 1) converting each earnings call transcript (”gross”, not preprocessed) into a set of sentences rather than treating the entire transcript as a single document. This allows the model to detect more granular signals about trade-related sentiment within specific parts of the text.

Because the transcripts are initially unlabeled, 2) we use a zero-shot classification model ("facebook/bart-large-mnli") to assign each sentence a preliminary sentiment label: positive, negative, or neutral. To ensure reliability, only sentences with a class probability above a chosen threshold are kept; other sentences remain unlabeled. This filtering step improves the quality of the training data by focusing on sentences that the model can classify with high confidence.

To choose the optimal threshold, we manually labeled 50 sentences. Then, we compute the accuracy between machine (zero shot) and human classification over the 50 sentences and by varying the threshold in [0.33, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]. We then choose the threshold giving the maximum accuracy.

Once the sentences are labeled, 3) we build a TF-IDF representation of the textual data (to represent them numerically). This representation assigns higher weights to words that are frequent within a sentence but less common across all sentences, emphasizing terms that carry more discriminative information.

Using this TF-IDF matrix as input, 4) we train a logistic regression classifier to predict the sentiment label of new sentences. The model is trained on a split dataset (training, validation, and test sets) to ensure proper evaluation and avoid overfitting. On the validation set, the model achieves c. 80% accuracy.

Finally, 5) the sentence-level predictions are aggregated back to the transcript level using majority voting of predicted labels. This transcript-level classification is then used to group firms into "trade peace" (positive class) and "trade war" (negative class) categories.

The predictive ability of this model is assessed via an event-study analysis around major U.S. tariff announcements. If the model successfully captures firm-level trade sensitivity, we expect to observe statistically significant differences in abnormal returns between firms classified as trade peace and trade war during these events.

4.4 A transformer approach: FinBERT

We use FinBERT to analyze the sentiment at the sentence-level if the sentence contains a trade related keywords. The FinBERT model then classify each sentence as 'positive', 'negative' or 'neutral' with an associated confidence score. Then we compute the 'trade exposure' at the transcript level by differencing the average negative confidence score and average positive confidence score of each trade related sentences in the transcript. So, a negative trade exposure score means that there is more positive trade related sentences than negative ones within a given transcript. Accordingly, we classified trade peace stocks as the firms with negative trade exposure score, which we expect to perform better on trade policy shock events. Conversely, trade war stocks are the firms with a positive trade exposure score.

We note that to avoid forward looking bias, we'll look at the results of the FinBert model for event dates after August 2019, release date of the research paper about FinBERT.

We report below the number of TP&TW stocks at the quarter preceding event dates below.

Quarter	TP	TW
2017-01-01	132	36
2017-04-01	39	15
2017-10-01	24	16
2018-01-01	52	10
2018-04-01	124	30
2019-01-01	180	53
2019-04-01	176	50
2019-07-01	202	65
2019-10-01	179	56
2024-01-01	40	10
2024-10-01	148	24
2025-04-01	724	93

Table 5: Number of TP&TW stocks at the quarter preceding event dates.

5 Assessing our measures: an event-study of stock prices around trade policy shocks

5.1 Event-Study methodology

To test empirically whether our measures capture exposure/is being able to classify stocks to shifts in trade policy, we conduct an event study of reactions to trade policy events. Focusing on the period following the selected events allows us to link changes in market expectations to the common narrative that characterizes the events, which in our case is a shift in trade policy. This means that price reactions to policy shifts represent an ex-post measure that captures market participants' assessment of exposure of firms to the policy shift. In practical terms, given that our hypothesis is that trade peace stocks stand to lose less than trade war stocks when governments raise barriers to international trade, we test whether the stock price reactions of trade peace stocks to these events are significantly different than those of trade war stocks.

First, we define event days on which there is a clear change in trade policy. To do that we use the dates provided by Scientific Beta supplemented by ChatGPT recommendations for after 23 August 2019 dates.

We also need to specify the event window in which we expect the event to impact stock returns. The event dates define the moment when the news of a trade policy change reaches the market. Only analysing returns on this day, however, may not allow enough time for the market to fully adjust prices to the event. The event window should also not be too long since we want to avoid capturing price reactions to subsequent non-trade related news. Therefore, we define the main event window to include the event date and the subsequent six calendar days. The length of these event windows is in a similar range as those used by Davies and Studnicka (2018) and Amiti, Kong, and Weinstein (2020).

Then, we want to estimate the stock-specific impact of the events on the share price during the event windows. To isolate the stock-specific component of the stock returns, it is necessary to control for the factors driving returns independent of the event. Therefore, we estimate a model for each stock's expected returns based on daily returns from one year before each event until one week before the event[1]. Our main results are based on the classic CAPM as the model for normal returns.

$$R^i_t = \alpha^i + \beta^i * MKT_t + \varepsilon^i_t \tag{3}$$

R^i_t represents the returns of stock i at day t in excess of the risk-free rate and MKT_t is the return on the market factor.

For each day t in the event windows, we can then define the abnormal return AR^i_t for stock i as the difference between the realised return RR^i_t and the normal expected return NR^i_t. In the equations below, RF_t is the risk-free rate on day t and $\hat{\beta}^i$ is the estimated value of β^i above.

$$AR^i_t = RR^i_t - NR^i_t, \quad \text{with} \quad NR^i_t = RF_t + \hat{\beta}^i * MKT_t. \tag{4}$$

Finally, the cumulative abnormal return CAR^i for stock i, over a given event window, is the sum of its abnormal returns on each day in that window[2]. Since the event windows are chosen as moments on which shifts in trade policy are expected to influence stock prices and market-wide returns are controlled for, the cumulative abnormal returns capture the stock-specific impact of trade policy shifts. Therefore, we interpret it as an ex-post measure of exposure to such shifts.

$$CAR^i = \sum_{t \in \text{event window}} AR^i_t. \tag{5}$$

We expect the cross-sectional mean of CAR to be different between trade peace and trade war stocks if our measures are relevant.

[1] If the period from one year until one week before an event includes other event windows, we exclude these event windows from the data when estimating the model for normal returns.

[2] We follow standard practice in event study and take the sum of daily returns as opposed to the product of cumulative daily returns. This allows us to conduct standard hypothesis for arithmetic averages (i.e., a t-test). Given the short time span of events (a few days), the difference between the two ways of computing cumulative returns over the event window will be tiny.

5.2 Returns-based validation

5.2.1 BoW

We present below the cumulative abnormal returns (CAR) averaged cross-sectionally (across stocks) around each event date (7 days window including the day of the event).

Table 6: Cross-sectional Mean CAR for Trade Peace and Trade War Firms

Date	Trade Peace	Trade War
2017-04-24	-0.001611	-0.005896
2017-08-08	0.001522	-0.004264
2018-01-01	-0.006875	-0.007782
2018-03-01	-0.003626	0.001751
2018-03-22	0.003095	0.002686
2018-04-02	0.000518	-0.000002
2018-06-15	0.001455	0.003223
2018-09-17	-0.006555	-0.007920
2019-05-10	-0.020097	-0.003474
2019-08-23	0.000471	-0.003659
2019-09-01	0.039224	0.011657
2019-12-15	-0.002745	0.001544
2020-01-15	-0.013681	-0.008064
2024-05-14	-0.027461	-0.020215
2025-02-01	-0.012865	-0.006314
2025-02-04	-0.018939	-0.010065
2025-03-04	-0.006978	-0.006479
2025-03-12	-0.014697	-0.000275
2025-07-23	NaN	NaN
Mean	**-0.499126**	**-0.353042**

The difference in CAR between trade peace and trade war stocks is -0.15% and the frequency of positive difference in CAR between trade peace and trade war stocks is 44.44%. We remark that the mean CAR of trade peace stocks is more negative that the one for trade war stocks signaling that our hypothesis that firms mentioning trade related words are more informed and better prepared for trade policy shocks contrary to firms that did not mention any trade keywords is false. To rigorously and clearly conclude, we'll need to assess the statistical significance of the difference in mean CAR between TP&TW stocks but as "normal" t-stats are not optimal [3] on event-study analysis, we will conclude without computing the statistical significance (this is only due to time concerns).

5.2.2 BoWws

We present below the cumulative abnormal returns (CAR) averaged cross-sectionally (across stocks) around each event date (7 days window including the day of the event).

[3] The t-statistics must be adjusted for event-induced variance and correlation across stocks, following Kolari and Pynnonen (2010)

Table 7: Cross-sectional Mean CAR for Trade Peace and Trade War Firms

Date	Trade Peace	Trade War
2017-04-24	-0.001431	0.003721
2017-08-08	-0.003429	0.016194
2018-01-01	0.012747	-0.003188
2018-03-01	0.010507	0.003635
2018-03-22	-0.057545	-0.003456
2018-04-02	0.001049	-0.014346
2018-06-15	0.006528	-0.006257
2018-09-17	-0.010983	-0.000636
2019-05-10	-0.020627	-0.023910
2019-08-23	0.002633	-0.001940
2019-09-01	0.036689	0.040992
2019-12-15	-0.007257	-0.005209
2020-01-15	-0.015615	-0.012599
2024-05-14	-0.054909	-0.018841
2025-02-01	-0.015376	-0.013584
2025-02-04	-0.012886	-0.023057
2025-03-04	-0.024500	-0.001109
2025-03-12	-0.023214	-0.017392
2025-07-23	NaN	NaN
Mean	**-0.986767**	**-0.449908**

The difference in mean CAR between trade peace and trade war stocks is -0.54%. The frequency of positive difference in CAR between trade peace and trade war stocks is 38.89%.

We remark that mean CAR is more negative for trade peace stocks than for trade war ones. This indicates either that our BoWws measure does not capture the right sentiment around trade keywords or that firms talking positively about trade keywords are the ones that perform poorly in the stock market on trade event dates.

5.2.3 TF-IDF - logistic regression

We present below the cumulative abnormal returns (CAR) averaged cross-sectionally (across stocks) around each event date (7 days window including the day of the event). We note that we have a lots of nan because for this model we need training and validation data. The sample is split into three periods: a training period from May 10, 2007 to January 1, 2019, a validation period from January 2, 2019 to January 1, 2024, and a test period from January 2, 2024 to July 24, 2025. The test period corresponds to the period where we perform the event study analysis.

Table 8: Cross-sectional Mean CAR for Trade Peace and Trade War Firms

Date	Trade Peace	Trade War
2017-04-24	NaN	NaN
2017-08-08	NaN	NaN
2018-01-01	NaN	NaN
2018-03-01	NaN	NaN
2018-03-22	NaN	NaN
2018-04-02	NaN	NaN
2018-06-15	NaN	NaN
2018-09-17	NaN	NaN
2019-05-10	NaN	NaN
2019-08-23	NaN	NaN
2019-09-01	NaN	NaN
2019-12-15	NaN	NaN
2020-01-15	NaN	NaN
2024-05-14	-0.007923	-0.037914
2025-02-01	-0.014883	-0.012699
2025-02-04	-0.025851	-0.019302
2025-03-04	0.031740	-0.004524
2025-03-12	-0.007351	-0.014419
2025-07-23	NaN	NaN
Mean	**-0.485353**	**-1.777159**

The difference in CAR between trade peace and trade war stocks is 1.29% and the frequency of positive difference in CAR between trade peace and trade war stocks is 60.00%. We remark that for this model, the mean CAR is less negative for the trade war stocks than for the trade peace ones, whereas it was the inverse for the BoW and BoWws models. This indicates that firms mentioning a majority of positive sentences about trade among all the sentences related to trade keywords are likely to perform better in the stock market but perform still negatively. We do not managed to get different mean CAR sign.

5.2.4 FinBERT

We present below the cumulative abnormal returns (CAR) averaged cross-sectionally (across stocks) around each event date (7 days window including the day of the event).

Table 9: Cross-sectional Mean CAR for Trade Peace and Trade War Firms

Date	Trade Peace	Trade War
2017-04-24	NaN	NaN
2017-08-08	NaN	NaN
2018-01-01	NaN	NaN
2018-03-01	NaN	NaN
2018-03-22	NaN	NaN
2018-04-02	NaN	NaN
2018-06-15	NaN	NaN
2018-09-17	NaN	NaN
2019-05-10	NaN	NaN
2019-08-23	NaN	NaN
2019-09-01	0.038412	0.038884
2019-12-15	-0.003160	0.000042
2020-01-15	-0.016415	-0.014867
2024-05-14	-0.032083	-0.021344
2025-02-01	-0.014932	-0.007586
2025-02-04	-0.020682	-0.016934
2025-03-04	-0.008770	0.025794
2025-03-12	-0.015301	-0.007902
2025-07-23	NaN	NaN
Mean	**-0.911641**	**-0.048929**

The difference in CARs between trade peace and trade war stocks is -0.86% and the frequency of positive difference in CAR between trade peace and trade war stocks is 0.00%. We remark that mean CAR is more negative for trade peace stocks than for trade war ones. This indicates either that our BoWws measure does not capture the right sentiment around trade keywords or that firms talking positively about trade keywords are the ones that perform poorly in the stock market on trade event dates.

5.3 Models comparison

In this subsection, we present the same results as above but unified on the available dates for all models to ease comparison.

Table 10: Summary table of cross-sectional mean CAR trade peace/war by model and event date

	bow		bowws		cm		fb	
	trade peace	trade war	trade peace	trade war	trade peace	trade war	trade peace	trade war
2024-05-14	-0.02746	-0.02022	-0.05491	-0.01884	-0.00792	-0.03791	-0.03208	-0.02134
2025-02-01	-0.01287	-0.00631	-0.01538	-0.01358	-0.01488	-0.01270	-0.01493	-0.00759
2025-02-04	-0.01894	-0.01006	-0.01289	-0.02306	-0.02585	-0.01930	-0.02068	-0.01693
2025-03-04	-0.00698	-0.00648	-0.02450	-0.00111	0.03174	-0.00452	-0.00877	0.02579
2025-03-12	-0.01470	-0.00027	-0.02321	-0.01739	-0.00735	-0.01442	-0.01530	-0.00790
Mean	-1.618794	-0.866948	-2.617712	-1.479684	-0.485353	-1.777159	-1.835372	-0.559453

The time-series difference mean of CAR between trade peace and trade war stocks for the Bow Model is -0.75%, for the BoWws -1.13%, for the cm (TF-IDF - logistic regression) 1.29% and for the fb (FinBERT) model -1.27%. The frequency of positive difference in CAR between trade peace and trade war stocks for BoW model is 0.00%, for BoWws model 20.00%, for cm model 60.00% and for the fb model 0.00%.

We thus conclude that our TF-IDF - logistic regression model is the best model in terms of discrimination ability between trade peace and trade war stocks and their returns (because this is the model with the highest absolute time-series difference mean CAR between trade peace and trade war stocks).

6 Conclusion

6.1 Summary of findings

In this study, the goal was to assess the classification ability of different NLP models using earnings call transcripts of the Russell 1000 constituents from 2007 to 2025. More precisely, we wanted to classify stocks either as trade peace (benefiting from trade policy shocks) or trade war stocks. We measured the performance of each model by computing the difference time-series mean of cross-sectional mean CARs (cumulative abnormal returns) between trade peace and trade war stocks during policy shocks events.

Our BoW model classified stocks as trade peace if the firms mentioned at least one trade related keywords in their earnings call preceding the event dates and as trade war otherwise. The BoWws model, which is a refinement of the BoW model to capture sentiment around trade related keywords classified as trade peace the firms with positive sentiment around trade keywords and as trade war otherwise. Finally, the TF-IDF - logistic regression model classified as trade peace the stocks with more positive sentences about trade than negative sentences in their earnings call and as trade war otherwise.

We found that (even if the statistical significance must be computed to rigorously and clearly conclude) none of the models is able to discriminate between trade peace and trade war stocks such that their difference time-series means of cross-sectional mean CARs between trade peace and trade war stocks is of opposite sign but we also found that the TF-IDF - logistic regression model is the one which leads to the highest absolute difference of time-series mean of cross-sectional mean CARs between trade peace and trade war stocks. So, in this sense, we can say that this is the best model.

6.2 Limitations

One of the limitations of our work is that we need to compute the statistical significance to rigorously and clearly conclude on our results by correcting the t-stats as recommended in Kolari and Pynnonen (2010) because the t-statistics need to be adjusted for event-induced variance and correlation across stocks.

Another limitation is that for some events, our sample size (number of stocks with measures available) is small, hence the above requirements. Also, we compare models but the selected stocks as trade peace and trade war are not exactly the same across models. Thus, performance difference may also come from sample differences rather than pure improvement.

Additionally, we have limited sample size for TF-IDF model (2024-2025 test period) as trade policy shocks are rare.

Lastly, we obtained mixed directional results across models.

6.3 Future work

Next steps are: 1) trying different models (LLMs such as GPT for example, or by changing the classification model of the TF-IDF - logistic regression model by a random forest), 2) trying other vectorization than the TF-IDF to encode text into numerical values and 3) perform the same study but using 10-Ks instead of earning calls transcripts (in progress).

7 Appendices

7.1 TF-IDF – Logistic Regression Model

The TF-IDF – Logistic Regression model combines two components: **(i) TF-IDF vectorization**, which converts textual data into numerical features by capturing the relative importance of words, and **(ii) Multinomial Logistic Regression**, a linear classifier that estimates class probabilities for discrete outcomes.

TF-IDF Vectorization. Term Frequency–Inverse Document Frequency (TF-IDF) is a weighting scheme that reflects how important a word is to a document relative to the entire corpus. For each document d and term t, the TF-IDF weight is computed as:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t)$$

where:

- $\text{TF}(t,d) = \frac{\text{count of term } t \text{ in } d}{\text{total terms in } d}$ is the term frequency,

- $\text{IDF}(t) = \log \frac{N}{1+n_t}$ is the inverse document frequency, with N the total number of documents and n_t the number of documents containing t.

In our pipeline:

1. We provided to the TF-IDF function all sentences ('gross', not preprocessed) containing trade related keywords from earnings call transcripts.
2. A TF-IDF matrix of size (number of transcripts) × (vocabulary size) was constructed.
3. This matrix served as the input for the logistic regression classifier.

Multinomial Logistic Regression. Logistic regression models the probability that an observation belongs to a certain class. For a given feature vector $\mathbf{x}$, the probability of class k is:

$$P(y = k|\mathbf{x}) = \frac{\exp(\mathbf{w}_k^\top \mathbf{x} + b_k)}{\sum_j \exp(\mathbf{w}_j^\top \mathbf{x} + b_j)}$$

where $\mathbf{w}_k$ and b_k are the parameters for class k.

Training procedure:

- We used a multinomial loss function (cross-entropy) optimized with the `lbfgs` solver.
- L2 regularization was applied to prevent overfitting due to the high-dimensional feature space.
- Hyperparameters such as the regularization strength C was set to the default value (1.0), max_iter=1000 and class_weight="balanced" (as the vast majority (65%) of sentences were classified as negative).

Application to trade-sensitivity classification.

1. The TF-IDF features served as input to train the logistic regression classifier.
2. Transcripts were labeled as **positive**, **negative** or **neutral** based on majority voting of label for all trade related sentences.
3. **Trade peace** stocks are the firms with **positive** transcript and **trade war** stocks otherwise.

Results in our study. The TF-IDF – Logistic Regression model achieved the largest absolute difference in time-series mean of cross-sectional mean cumulative abnormal returns (CARs) between trade peace and trade war stocks, with a difference of **1.29%**. Although this does not establish strong discriminative power (statistical significance remains to be assessed), this approach outperformed Bag-of-Words and FinBERT in our event study evaluation.

7.2 FinBERT

FinBERT is a domain-specific language model based on the BERT (Bidirectional Encoder Representations from Transformers) architecture, fine-tuned for financial text analysis. It was developed to address the limitations of general-purpose language models in capturing the nuances of financial terminology and sentiment.

Model architecture. FinBERT is built on top of BERT-base, which consists of:

- 12 transformer encoder layers,
- 12 self-attention heads,
- 110 million parameters,
- a hidden size of 768.

The model benefits from BERT's bidirectional attention mechanism, which allows it to learn context from both directions of the input text.

Domain-specific pretraining. FinBERT was further pretrained on large corpora of financial text (e.g., earnings call transcripts, financial news, analyst reports). This domain adaptation enables the model to better understand finance-related vocabulary, context, and sentiment nuances compared to generic BERT.

Sentiment classification. FinBERT was fine-tuned on the Financial PhraseBank dataset, which contains manually labeled financial sentences with positive, neutral, or negative sentiment. The fine-tuned version is widely used for:

- Sentiment analysis of financial documents,
- Event-driven trading signals,
- Risk assessment based on qualitative disclosures.

Advantages and limitations.

- **Advantages:**
  - Pretrained specifically for the financial domain, leading to better understanding of finance-related language.
  - Outperforms generic models in tasks such as sentiment classification for financial texts.
  - No need for manual feature engineering.
- **Limitations:**
  - The model is trained for sentiment classification, not directly for trade-sensitivity classification.
  - Computationally expensive for large-scale inference on tens of thousands of transcripts.
  - May miss subtle trade-related contexts that are not explicitly linked to financial sentiment.

Relevance to our pipeline. FinBERT was integrated as one of the classification approaches to assess firms' trade exposure. By leveraging its domain-specific sentiment analysis, we aimed to identify firms that discussed trade topics positively or negatively in their earnings calls.

7.3 Zero-Shot Classification and the `bart-large-mnli` Model

Zero-shot classification refers to the task of assigning labels to a given text without having trained a supervised classifier on examples of those labels. Instead, a pre-trained natural language inference (NLI) model is leveraged to determine whether a text is *entailment*, *neutral*, or *contradiction* with respect to a given hypothesis (i.e., a potential label expressed as a natural language sentence).

Underlying principle. In the zero-shot setting, the classification problem is reformulated as an NLI task. For each candidate label (e.g., "This text is about trade policy"), the model computes the probability that the input text *entails* the hypothesis. The label with the highest entailment probability is selected. This approach enables classification into arbitrary labels without fine-tuning.

The `bart-large-mnli` model. The model used is `facebook/bart-large-mnli`, a transformer-based sequence-to-sequence model trained on the Multi-Genre Natural Language Inference (MNLI) dataset.

- It has 24 layers, 16 attention heads, and 406M parameters.
- The model was trained on 433k examples covering a variety of genres and topics, which makes it effective for general-purpose inference tasks.
- The large architecture allows it to capture nuanced semantic relationships between texts and hypotheses.

Application to our study. We applied zero-shot classification using `bart-large-mnli` to assign trade-related labels to earnings call transcripts without requiring labeled training data. Specifically:

1. Each trade-related sentence of the transcripts was fed to the model along with candidate hypothesis: *"This sentence is positive"*
2. We fed three candidate labels: 'positive', 'negative' and 'neutral'.
3. The model computed the entailment score for candidate label.
4. If the maximum score among the 3 labels exceeded a defined threshold (0.8 in our case), the sentence was classified according to its label with the highest score; otherwise the sentence was classified as 'unlabeled'.

Advantages and limitations.

- **Advantages:**
  - No labeled dataset is required for training.
  - Flexible labeling: new categories can be tested by simply changing the hypotheses.
  - Strong generalization due to the large pre-training corpus.
- **Limitations:**
  - Performance depends on the phrasing of the hypotheses.
  - Computationally more expensive than simple bag-of-words approaches for large corpora.

References

[1] Giovanni Bruno, Felix Goltz, Ben Luyten (2023). Firm-Level Exposure to Trade Policy Shocks: A Multi-dimensional Measurement Approach. *European Financial Management*, Volume 30, Issue 4, September 2024.

[2] James W. Kolari, Seppo Pynnönen (2010). Event study testing with cross-sectional correlation of abnormal returns. *The Review of Financial Studies*, Volume 23 Issue 11 November 2010.