Examine the correlation between Twitter Sentiment data and the stock data for 4 big tech companies

Anonymous ACL submission

Abstract

In the experiment examining the impact of tweets on the stock data of the four major companies—Apple, Amazon, Google, and Microsoft—we employed two sentiment analysis models. These are used to analyse the sentiment of all tweets, that are associated with the company, to get positive, neutral and negative daily percentages. In detail we can than analyse if these percentages on one day makes a difference for the stock data on the next day. We use the Pearson coefficient, as well as the euclidean distance and a MSE calculation to analyse the datasets. In general the Apple dataset is best predictable and the Google dataset performs worst, but it has to be considered that the evidence of this correlating effect is rather low.

1 Introduction

The main focus of this project is to use sentiment analysis on twitter. The aim is to search for patterns or correlation with the stock data of four major tech giants: Amazon, Apple, Google, and Microsoft. We will check if positive, negative and neutral percentages of daily twitter posts have a lagging correlation of one day.

To address this inquiry we will deploy diverse methods like calculating the Pearson coefficient for the Euclidean distances of Twitter posts on individual days and assessing the predictiveness of future values. This process will be carried out using two models: a general sentiment model and a social media specific sentiment model.

This analysis holds the potential to provide insights into the predictability of specific companies' stocks and their distinctions from one another. Should the findings not align with expectations, future analyses can explore alternative datasets beyond Twitter or consider different time frames than those utilized in this study.

2 Theory

In this section theories and methods that underpin the project and experiments will be elaborated.

2.1 Sentiment Analysis

In the context of text mining, sentiment analysis is used to identify the mood within a given text. Instead of manually sifting through diverse sources like reviews, comments, or tweets, sentiment analysis tools offer an automated approach to determine the emotional tone of a text.

Numerous pretrained models are readily available on the Hugging Face hub : (Delangue, 2016), each tailored for specific use cases. As one main goal of this analysis is to calculate the sentiment of specific tweets, the usage of the standard sentiment model which is based on the Distilbert-base model (Sanh et al., 2020) as well as the bertweet model(Pérez et al., 2021a), optimized for tweets, is used.(Pascual, 2022)

2.1.1 Distilbert-base model

The standard model for our experiments is a Distilbert-base model, which is a distilled version of the BERT base model being 40% smaller than the original, but still retaining 97% of its language understanding. Therefore the computation time could also be reduced by 60%.(Sanh et al., 2020) The original model was trained on books and english wikipedia entries. For fine tuning the classification of the model for the sentiment analysis process, the sst-2 dataset was used.(Face, 2019) This dataset comprises 67,300 fully labeled entries sourced from movie reviews categorized as either positive or negative .(Socher et al., 2013)

2.1.2 Bertweet-base-sentiment-analysis

The foundational model for Bertweet-base-sentiment-analysis is the BERTweet model, sharing the architecture with the original BERT model and adopting the training methodology of the RoBERTa

and XLM-R models. Given the well-established nature of the BERT model and RoBERTa training procedure, a detailed description is omitted but can be explored further in references (Devlin et al., 2019), (Zhuang et al., 2021). The training of the BERTweet model is done based on 845 million English tweets, gathered in the years 2012-2019 and additional 5 million tweets related to the Covid-19 pandemic. Retweeted tweets were filtered and emojis were translated to text icons. For tweet length considerations, only tweets with a token count between 10 and 64 were included (Nguyen et al., 2020). In the specific context of sentiment analysis, the BERTweet model underwent further refinement and training for classifying tweets into positive, negative, and neutral sentiments, resulting in the Bertweet-base-sentiment-analysis (Pérez et al., 2021b). The training dataset consisted of 100,000 English entries, individually labeled on a scale from (-2) to 2, representing the states: STRONGLYNEGATIVE, WEAKLYNEGATIVE, NEUTRAL, WEAKLYPOSITIVE, and STRONGLYPOSITIVE (Rosenthal et al., 2017).

2.2 Euclidean distance

In order to calculate similarity, the euclidean distance is a very popular measure. This measure can take multiple variables as input and can calculate the distance between these points. With this single measure different variables of one entry can be combined and can be compared to other entries, offering a means to evaluate the similarity between them. For the 3 dimensional case it would look like the following (Schichl and Steinbauer, 2018).

$$distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

2.2.1 Pearson correlation

The Pearson correlation emerges as a highly regarded measure that quantifies the strength of a linear relationship between two variables, denoted as x and y. A coefficient nearing 1 indicates a perfect linear correlation, while proximity to 0 signifies a weaker linear relationship between the variables. It's essential to note that a coefficient close to 0 does not imply the absence of correlation; rather, it suggests a lack of linear correlation. The coefficient is defined as the cross-correlation between x and y divided by the variances of x and y, as illustrated below (Benesty et al., 2008).

$$\rho = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

In order to check for significance in the Pearson correlation, a t-test for regression models is used, with the following formula:

$$\frac{\beta}{se/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Here β represents the coefficients of the input values, "se" represents the standard error and the term beneath the square root represents the sum of squares between the mean and x.

2.3 Linear regression model

Assuming a model Y dependent on x, which assumes a underlying linear function. With this general assumption we can build a linear regression model.

$$Y = \beta X + \epsilon$$

Here β denotes a vector of regression coefficients. In order to calculate the perfect β for the underlying data, we use a minimizing algorithm for beta to reduce the error of new predicted values. (Magnus and Magnus, 2019)

To get an estimate how predictive the values for the model are, assessing the Mean Squared Error (MSE) is a conventional approach. It measures the error how far the predicted values are away from the correct ones on average. The calculation is expressed by the formula. (Pishro-Nik, 2014)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

3 Data

The dataset employed for sentiment analysis and its subsequent correlation with stock data is the "Tweets about the Top Companies from 2015 to 2020" (Ömer Metin and Dogan, 2020) dataset for tweets, complemented by the "Stock Market Dataset" (Onyshchak, 2020). Our analysis exclusively focuses on data from the year 2015, as further exploration of subsequent years makes it impractical to yield meaningful output within the designated time frame.

The dataset of twitter entries consists of over 3 million twitter entries related to Apple, Amazon, Google, Microsoft and Tesla. These entries consist of tweets where the mentioned companies are tagged, featuring key information such as company details, tweet_id, writer, post date, body, comment number, retweet number, and like number. For the

	Complete Dataset	2015
Apple	1.425.013	360.177
Amazon	2.143.728 2	210.700
Google	1.817.582	475.698
Microsoft	1.800.724	417.148

Table 1: Tweets per company

analysis of the sentiments and finding a possible correlation with the stock data, the focus will lie on the body (text of specific post) and the company. The amount of tweets per company can be seen in table 1

The second dataset, housing stock data for the companies, comprehensively records daily stock information, including Open, High, Low, Close, Volume, and Adjusted Close In our experiment we will focus on the Adjusted Close data, as this includes the corporate actions like dividends, stock splits at the end of a day as well.

Our analysis spans 252 days in the year 2015, aligning with the number of stock trading days. Weekends, not considered as trading days, are thus excluded from the dataset.

4 Results

In the following section the results for the experiments will be represented. We perform calculations to determine the percentage of positive, neutral, and negative posts each day, comparing these metrics with the stock market difference from the preceding day. With that information three graphs per model and company stock data are drawn. Additionally the euclidean distance is calculated for the positive, neutral and negative percentages as one value. The Pearson coefficient is also calculated for the euclidean distance and for positive, neutral and negative percentages individually. Lastly it has to mentioned, that the graphs for stock data were normalized between 0 and 1 and the neutral values in the normal model always shows the same value 0 as the normal model is not capable of classifying posts as neutral.

4.1 Apple

As illustrated in Figure 1 and 2, both the normal and social media models for Apple exhibit a positive trend in the percentage of positive values and a negative trend in the percentage of negative values. The neutral graph for the social media model maintains a relatively constant slope with only a slight incline. The same tendencies can also be seen in the Pearson coefficient values in table 2 and 3 as a slight positive correlation can be seen for positive percentages and a slight negative correlation can be seen for the negative percentages. The p-value 0.0000 is not equal to zero, but the value is too small to show any output with 4 decimal numbers. Lastly for checking predictability, a computation of the MSE for the linear regression and a computation of the MSE with the mean of all values is done. The second value should build a baseline for the linear regression MSE. This is done for the other companies as well. The values for both models and both MSE prediction can be seen in the tables 7, 9, 8, 10. As visible the MSE values for the linear regression are smaller for the Apple dataset for all variables. Run times to do the sentiment analysis for the percentages can be seen in table 4.

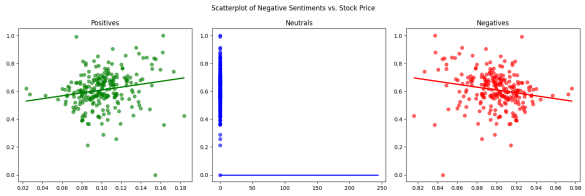

Figure 1: Apple normal model

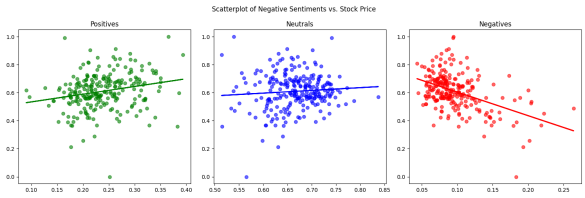

Figure 2: Apple social media model

4.2 Amazon

The graph results of the Amazon data set show similar tendencies as for the Apple data set, but less pronounced. While the normal model displays reasonable slopes for positive and negative values, the social media model's positive slope, with a Pearson coefficient of 0.0664, closely approaches the neutral slope, differing by only around 0.02. Like in the Apple dataset, the p-values for the Amazon dataset are remarkably low, except for the neutral values in the social media model. Although the MSE values for the Amazon dataset are lower across all variables compared to the base MSE, the difference is not as substantial as observed in the Apple dataset. The run time for the Amazon model is dramatically longer (a little below one and a half hours longer

than the Apple dataset), which is due to the amount of processed tweets in the data set.

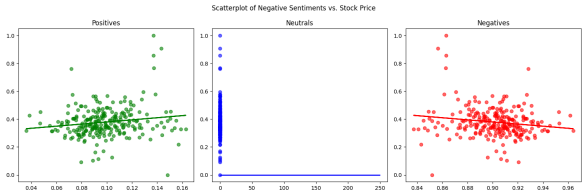

Figure 3: Amazon normal model

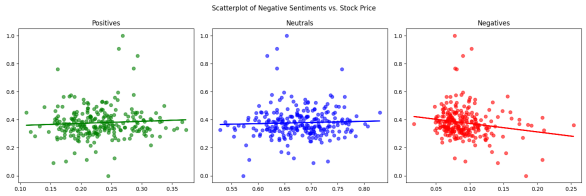

Figure 4: Amazon social media model

4.3 Google

The tendencies in the Google dataset reveal two notable outputs: the Pearson coefficient for positive percentages and the Pearson coefficient for negative percentages. The other Pearson coefficients are all below 0.1 and therefore the graphical illustration does not show a instantly interpretable tendency as well. Worth mentioning is also the Pearson coefficient for the euclidean distance having a value close to 0. The p-value shows a high value for the neutral and euclidean distance. The MSE values for the linear regression are not significantly bigger. Especially the positive mean MSE is out of the line compared to the other values.

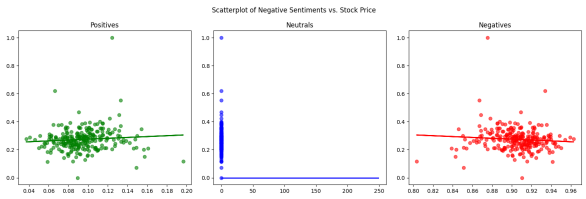

Figure 5: Google normal model

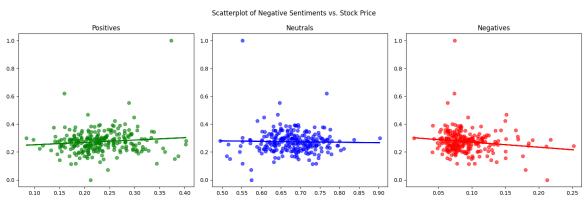

Figure 6: Google social media model

4.4 Microsoft

Lastly, for the Microsoft dataset, the social media model stands out as the only model with a steeper slope than one of the individual percentages (positive, negative), surpassing the positive slope by 0.06. The tendencies with the normal model are below 0.1 as well and therefore do not have a real tendency towards any direction. Unlike the Google model, the Pearson coefficient for the Euclidean distance in the social media model is higher and closely approaches the value observed in the Apple dataset. The mean MSE for the Microsoft dataset quite close to the linear regression MSE except in the normal model.

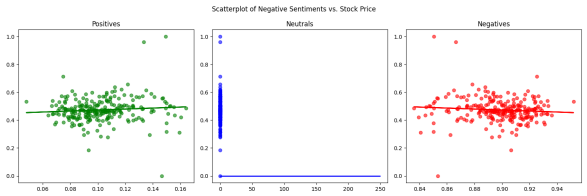

Figure 7: Microsoft normal model

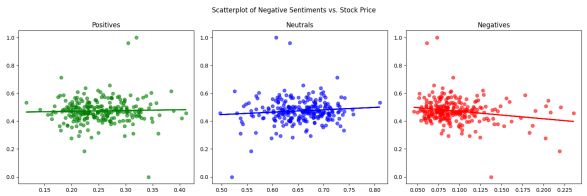

Figure 8: Microsoft social media model

5 Discussion

In the following section the results for the individual companies are discussed and compared to each other. Additionally limitations occurring in the project will be discussed.

5.1 Apple

Examining the graphs for Apple, a discernible upward slope is observed in both the normal and social media model for the positive percentage, implying a potential relationship between positive percentages and stock values on the subsequent day. This inclination is further reflected in the Pearson coefficient, evident in Table 2 for the Apple company's positive percentage column. As we only classified between positive and negative values, the

Company	Positive	Negative	Together(Euclidean Distance)
Apple	0.1935	-0.1935	-0.1933
Amazon	0.1559	-0.1559	-0.1558
Google	0.0863	-0.0863	-0.0907
Microsoft	0.0878	-0.0878	-0.0879

Table 2: Pearson Coefficient normal model

Company	Positive	Neutral	Negative	Together(Euclidean Distance)
Apple	0.2264	0.0798	-0.4600	0.1591
Amazon	0.0664	0.0409	-0.1737	0.0577
Google	0.1113	-0.0240	-0.1419	-0.0012
Microsoft	0.0325	0.0980	-0.1997	0.1267

Table 3: Pearson Coefficient Social Media model

Company	Normal model	SM model
Apple	2,533	4,543
Amazon	3.384	6.034
Google	3.135	5.628
Microsoft	2.969	5.257

Table 4: Runtime in hours for the sentiment analzsis

negative coefficients are always the negative value of the positive value. This also needs to be considered for the analysis of the other companies.

The Euclidean distance exhibits a clear negative Pearson coefficient, primarily influenced by the prevalence of negative days around 90%, consequently leading to a dominant negative factor in the Euclidean distance calculation.

The values and graphs for the social media model look pretty much similar to the normal model, meaning a slight positive slope for the positive percentages and a slight negative percentage for negative stock difference. As expected, the neutral values exhibit no clear slope, with a Pearson coefficient closer to 0. Looking at the p-values, we do see significant values, except the neutral one for the social media model, with a value of 0.2131. As no positive or negative tendency was expected for this for the neutral tweet percentages, this p-value is also as expected.

The MSE values for the linear regression are much lower than the values we are getting if you only use the mean as a prediction. This means that for Apple the values are a little predictable for both models and all graphs.

5.2 Amazon

The graphs for the Amazon dataset, especially for the normal model, have a very similar output. The tendencies in the social media model are much flatter than for the Apple dataset, but slopes are still visible and the neutral graph is very flat. Consequently, the Pearson coefficients in table 2 and 3 show the same schema, of having less strong rising and falling numbers. Low p-values for the normal model and negative percentages in the social media model indicate a significant effect, while positive and neutral percentages exhibit p-values above 0.05, suggesting no clear trend.

Similar to the Apple dataset, the Amazon dataset seems to be a little predictive as well, especially the normal model. In the social media model the mean values show less disparity. However, with no statistical significance in the p-values for positive and neutral percentages, the predictability is not as pronounced.

5.3 Google

The Google dataset reveals even less pronounced positive and negative trends. As this cannot be directly visioned from the model graphs, the Pearson coefficients have to be considered to get a clear view on that. For most Pearson coefficients, the values hover around half of the slope observed in the Amazon dataset. This tendency is also shown in the p-values, as there is only the positive and negative percentages of the social media model, that seem to be significant. The Euclidean distance, not showing significance in any direction, may not be inherently negative, but given the marginal significance of positive and negative percentages, it suggests a lack of substantial impact.

Company	Positive	Negative	Together(Euclidean Distance)
Apple	0.0012	0.0012	0.0023
Amazon	0.0068	0.0068	0.0137
Google	0.0869	0.0869	0.1524
Microsoft	0.0832	0.0832	0.1659

Table 5: p-value of t-test

Company	Positive	Neutral	Negative	Together(Euclidean Distance)
Apple	0.0002	0.2132	0.0000	0.0127
Amazon	0.1479	0.5200	0.0029	0.3640
Google	0.0394	0.7059	0.0124	0.9854
Microsoft	0.3040	0.1222	0.0008	0.0453

Table 6: p-value of t-test

The google dataset seems to be not predictive as the mean values for calculating the MSE in the linear model are very close or even worse than just predicting the mean. As the values are very close together in the dataset (which can be seen in the graph), it is no wonder that the mean prediction is already a quite good one.

5.4 Microsoft

Lastly Microsoft, does not show real tendencies. Notably, even the regression line for positive percentages is flatter than the neutral model in the social media model, as reflected in the Pearson coefficients of 0.0325 and 0.0980. Only the negative percentage in the social media model displays a discernible negative slope, accompanied by a significant p-value. This negative tendency could be explained by the outliers, that are visible in the graph. The regression line in the social media model is dramatically influenced by the outliers towards the days of around 20 percent daily negative percentages.

A phenomenon akin to the Google dataset is apparent in the Microsoft dataset. The values, as depicted in the graph, are closely clustered, and the mean serves as a reasonably accurate prediction. Predictions based on the linear regression model are thus significantly constrained.

5.5 Limitation

Due to time constraints, we have limited our analysis to only one year (2015) from the entire dataset spanning from 2015 to 2020. While additional data would undoubtedly provide deeper insights for the study, the runtime for processing the current datasets alone already exceeded approximately 33 hours, making it impractical to include more data within the scope of this study.

Furthermore, we have solely utilized pre-trained models for sentiment analysis. Although the normal model, which is not improved on social media data extending the base model on Twitter data related to the analysed companies could have been an option, which could potentially yield to a more refined sentiment analysis model. Following this it has to be mentioned, that the sentiment analysis is not a 100% correct analysis as we are only predicting the sentiment and not having a correct value for the sentiments of the twitter posts. The effect of false sentiment predictions on the stock data, would therefore be interesting as well. Wrong influence the reliability of the graphs and the calculations.

Lastly it has to mentioned, that twitter data is often not the best to put the analysis on. Company related news are frequently disseminated through channels such as Bloomberg before it surfaces on Twitter. Consequently traders usually rely on Bloomberg as a source to estimate future sock market values. Due to the lack of availability of sources like that, twitter data was our go to data source for the analysis in this project.

6 Conclusion

While some graphs suggest subtle inclinations towards a positive or negative linear relationship between the percentage of positive or negative tweets, these tendencies are infrequently supported by p-values and Pearson coefficients. In general, for most graphs, these tendencies appear too faint given the dataset's volume, as explicitly illustrated in the p-value graphs for both the normal and social media models 5, 6.

Company	Positive	Negative	Together(Euclidean Distance)
Apple	0.0237	0.0237	0.0237
Amazon	0.0138	0.0138	0.0138
Google	0.0869	0.0869	0.0069
Microsoft	0.0116	0.0116	0.0116

Table 7: MSE values of the social media values

Company	Positive	Neutral	Negative	Together(Euclidean Distance)
Apple	0.0219	0.0232	0.0165	0.0225
Amazon	0.0140	0.0134	0.0127	0.0134
Google	0.0064	0.0064	0.0056	0.0062
Microsoft	0.0110	0.0102	0.0094	0.0099

Table 8: MSE values of the social media values

Upon closer inspection of values deemed significant, it becomes apparent that they primarily manifest in the negative section. This phenomenon seems to be a consequence of outliers distorting the slope, thereby subtly manipulating the p-value. Additionally we also see very high values for the p-value for the euclidean distances. This does not necessarily need to show that there is no tendency, but in combination with the individual positive, neutral and negative percentages this underlines the argument that the graphs do not have a real trend.

References

Jacob Benesty, Jingdong Chen, and Yiteng Huang. 2008. On the importance of the pearson correlation coefficient in noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4):757–765.

Clément Delangue. 2016. Hugging face. Accessed on December 27, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugging Face. 2019. Distilbert base uncased finetuned sst-2. Accessed on December 27, 2023.

Jan R. Magnus and Jan R. Magnus. 2019. *The linear regression model*. DOI.

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*, pages 9–14, Online. Association for Computational Linguistics.

Oleh Onyshchak. 2020. Stock market dataset.

Federico Pascual. 2022. Getting started with sentiment analysis using python. Accessed on 12 27, 2023.

Hossein Pishro-Nik. 2014. Mean squared error (mse). Accessed: December 31, 2023.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021a. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021b. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Hermann Schichl and Roland Steinbauer. 2018. *Einführung in das mathematische Arbeiten*. Springer Verlag, Heidelberger Platz 3, 14197 Berlin, Germany.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Company	Positive	Negative	Together(Euclidean Distance)
Apple	0.2860	0.1060	0.1094
Amazon	0.0896	0.2925	0.2985
Google	0.0332	0.4231	0.4299
Microsoft	0.1355	0.2054	0.2108

Table 9: MSE mean values of the social media values

Company	Positive	Neutral	Negative	Together(Euclidean Distance)
Apple	0.1649	0.0262	0.2882	0.0343
Amazon	0.0340	0.1055	0.0925	0.1366
Google	0.0065	0.1800	0.0338	0.2225
Microsoft	0.0514	0.0483	0.1397	0.0740

Table 10: MSE mean values of the social media values

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Ömer Metin and Mustafa Dogan. 2020. Tweets about the top companies from 2015 to 2020.