News Sentiment and Company Reports Impact on Stock Returns

Nikita Pashkov
HSE University
Moscow, Russia

Ilya Makarov
AIRI,
HSE University, ISP RAS
Moscow, Russia

2024 IEEE 28th International Conference on Intelligent Engineering Systems (INES) | 979-8-3503-6759-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/INES63318.2024.10629134

***Abstract*—Stock market returns prediction is a complicated topic with ever-growing interest within the past 20 years. Recent studies have been constantly trying to find more useful predictors, which quite often were found to be extracted from text sources such as social networks or discussion forums. In this study, we propose to take a closer look at the information contained in Securities and Exchange Committee (SEC) reports in the form of 10-K and 10-Q. In particular, we focus on analyzing the Management Analysis and Discussion (MD&A) section.**

To perform our analysis, we collected and processed almost 300,000 reports following the best practices of distributed computing on PySpark. It is worth noting that, unlike many other research studies in the same field, we consider not only S&P 500 companies but a broader range of companies as well. Based on the MD&A sections, we created a Bag-of-Words (BOW) corpus, and then utilized the TF-IDF technique to determine the weight of each word.

To evaluate the sentiment of the text, we utilized an exhaustive dictionary of words allocated to different emotional categories. This allowed us to compute sentiment scores for each category. Our findings indicate that certain emotions have a significant correlation with stock returns in the following quarter and can improve the performance of price movement direction classification models. Interestingly, commonly extracted *Positive* and *Negative* sentiment scores were found to have little importance, while the *Weak* and *Understatement* categories emerged as the top features.

***Index Terms*—Natural Language Processing, TF-IDF, Stock market**

I. Introduction

Both business and academic researchers have been trying to master the art of predicting stock price movements for a long time already. In the dawn, however, there were papers stating that this is an impossible task and that financial markets are efficient [1]. However, other recent studies appear to prove the opposite point. For example, [2] shows that predictions of Dow Jones Industrial Average (DJIA) are improved by some of the emotional categories of public mood, improving Mean Average Percentage Error (MAPE) of daily closing values by 6% and reaching over 86% up and down accuracy. The base of research being the Twitter feeds.

Another source of data which could potentially be useful in improving stock market predictions comes from Electronic Data Gathering, Analysis, and Retrieval (EDGAR) database, maintained by the United States Securities and Exchange Commission (SEC). This serves as a bridge to provide both corporate and individual investors with trustworthy data regarding the current state of businesses. In particular, firms are obliged to provide outlook on anticipated market risks, as well as to elaborate on corporate aims and means of achieving those. EDGAR currently offers a collection of over 150 distinct forms, although the most widespread ones are 8-K, 10-Q and 10-K.

Out of the above forms, 8-K is the closest to social networks in a way that it has no predefined schedule and can be published whenever firm has something important to say. 10-Q and 10-K, on the other hand, are filled periodically, are formalized strictly. The content is similar for the latter two, although the time coverage is of course different. Recent studies on these filings managed to establish significant correlation between the content and future firms' performance. Thus, [3] considered 8-K statements and their impact on stock prices within 3 days after publication. At the same time, (Qiu, 2007) investigated the effect of 10-K texts on one-to-two-year period after publication. To the best of this authors knowledge, there has not been a study covering both 10-Q and 10-K reports for all the companies traded in the US since year 2000 till recent days. Let alone the extraction of custom sentiment emotional categories (not a simple binary of positive or negative).

This paper focuses on parsing a huge volume of 10-Q and 10-K reports and extracting the MD&A section, which is a complex task (discussed in more details in the data chapter). The statements are later cleaned from unnecessary symbols, stop-words, etc. Remaining words are weighted in accordance with TF-IDF approach. Due to the dataset size, out-of-the-box solutions for the algorithm do not work, as it is not possible to fit the $documents \times words$ table in the memory. Author employs *PySpark* for distributed two-step computation of the scores. Harvard IV-4 General Inquirer [4] dictionary is used to infer mood categories of the words. Note that it is normal for one word to be related to several categories at the same time. *TF-IDF* weights are then summed to get the total score for each emotion. Because the length of MD&A sections is not always constant and is actually significantly less verbose in 10-Q compared with 10-K filings, the scores are normalized by the counts of symbols per MD&A. This assures that longer sections, which should naturally have more words of any emotions ceteris paribus, do not top the rankings of emotions content.

The target of the model is the stock price return starting from the filing publication date up until 1 day prior the publication of the subsequent filing. In case current filing is the

last (due to lack of data from EDGAR or delisting), 3 months are added to the publication date and the respective price is quoted. Default features are taken to be the last 4 quarters' returns. Higher numbers of lags were also tested, but did not provide a significant difference in performance and thus were omitted. Needless to say, the more lags are included in the dataset, the less companies satisfy the condition of having that much consequent filing quarters. Tuned Catboost model is fit to make predictions (with slight improvement over LGBM).

The paper presents several notable achievements. Firstly, both the sentiment augmented and default lags model outperform random predictions significantly. This demonstrates that sentiment in reports remains valuable even when making predictions over a relatively extensive time horizon. Secondly, it is evident that specific emotions carry more weight in predicting future stock price returns compared to the general *Positive* and *Negative* classes commonly associated with the term "sentiment." Concrete emotions prove to be more effective predictors.

From a data perspective, the paper introduces two significant developments:

- An MD&A section extraction script boasting an impressive accuracy rate of over 85
- Efficient distributed processing notebooks utilizing PySpark, capable of handling vast amounts of data.

The work is structured in the following way. First and foremost, literature review is carried to explore recent research in the field of sentiment analysis in application to predicting stock price returns. Consequently, data collection and preprocessing are covered, which includes working with historical stock prices, 10-Q & 10-K statements parsing, as well as MD&A extraction. Dataset is formed and cross-validated with events from the past. Later, sentiment categories are defined and elaborated. The approach for quantifying the emotional content is presented. Modeling goes in the next subsection, covering the target variable definition, training and testing procedures and the underlying algorithms used. Results follow in the next section. Finally, conclusions outline the achieved results and potential improvements to the current paper.

II. Related Work

Sentiment analysis is a popular method for evaluating attitudes, opinions, and emotions in social media and online platforms [5]–[10]. It has shown promise in predicting stock prices, which are influenced by economic indicators, company news, and global events.

Dohyun [11] studied the role of economic indicators in stock market fluctuations, finding no additional benefit of using macroeconomic data in predictions when historical volatility was considered. This contrasts with Paye [12], who established a clear connection between macroeconomic variables and stock volatility. Granger-causality testing confirmed the significance of variables such as commercial paper-to-Treasury spread, default return, and default spread in volatility predictions.

Ranco [13] and Bing [14] discovered a significant relation between Twitter sentiment and subsequent abnormal returns. This paper, however, focuses on sentiment analysis of fundamental news sources like 10-K and 10-Q financial statements, especially the Management Discussion and Analysis (MD&A) section.

The sentiment expressed in the MD&A section can provide valuable insights into a company's outlook and predict future stock returns. Feldman [15] found a significant correlation between extracted sentiment and stock returns using simple word frequencies. Davis and Angela [16], [17] found that earnings announcements are more positive than MD&A extracts.

Li [18] explored the readability of MD&A sections and found that poorer performing companies tend to write overly complicated reports. Bonsall [19] developed a readability measure and demonstrated a positive correlation between readability and firms with lower cost of debt and better ratings.

Altman's Z-scores [20], commonly enhanced with macroeconomic variables like in Andreeva [21], are widely used in research. Loughran and McDonald [22] developed their own dictionary based on SEC 10-K reports, finding that sentiment extracted from MD&A sections adds explanatory power to stock price return prediction models.

Gandhi [23] used sentiment as an independent variable to model financial distress in US banks, finding negative words in a 10-K report increased the probability of subsequent distress.

Sentiment analysis, particularly in MD&A sections of financial statements, has shown its usefulness in predicting stock price returns. Several studies explored the relationship between sentiment and stock market outcomes, highlighting the importance of textual data analysis in financial research.

One paper focused on extracting information from 10-Q reports is [24]. The authors generated TF-IDF weights for each document-word pair from S&P 500 companies' 10-Q statements. They aimed to predict "future stock price performance" using a target variable representing the difference between publication stock price and the stock's maximum/minimum value during a subsequent prediction period.

Another paper, [25], explored whether 10-Q and 10-K reports add value to the model. The study considered raw word count and TF-IDF weighting methods, using a sentiment index as a predictor for the future.

Yermack [26] examined the relationship between CEO's option awards and respective stock price fluctuations. Allen [27] used Thomson Reuters News Analytics (TRNA) articles to produce sentiment scores for companies traded in the DJIA index.

In summary, sentiment extraction from financial reports has been widely studied, with significant correlations and model performance improvements observed in predicting upcoming returns. Most studies focused on S&P 500 companies and did not explore alternative emotional categories [24]–[27].

III. Experiment Setting

A. Data Collection

First and foremost, the list of companies to work on is taken from a famous Sharadar [28] database via API. The main data asset of the paper is represented by SEC 10-Q & 10-K reports, which were parsed for the period spanning from the start of year 2000 up to end of 2022. The total number of parsed SEC reports is over 276.000 for 15.640 companies. Note that sometimes the SEC API returned errors, so not every report extraction is successful. But the errors are likely caused by extreme number of requests and thus do not possess a deterministic nature, which does not block from conducting the analysis. Stock prices parsing is way more trivial and is conveyed via *yfinance* [29] Python library, which conveniently returns prices ("Open", "High" and "Low", "Close", "Dividends" and "Volume") for every trading day. Note that the mandatory filings may in theory be published during a day with no trade, so in order to account for that the prices during the weekends are set to the latest available trading day values.

The most challenging data preprocessing step is the extraction of MD&A section. Open source did not feature any reasonable quality codes, so author proceeded with inventing the custom regular expression for parsing the aforementioned section. Notably, the SEC API provides two formats for working with the reports: *.html* and *.txt*, but after careful investigation it was established that *.html* format is bound to cover significantly less reports (i.e. in some cases *.txt* version of the report is present, while there is no *.html* analogue). After thoughtful consideration is was decided to proceed with working with richer dataset of *.txt* files. Despite the fact that *.html* may seem as an easier option in terms of consequent section parsing, this may not necessarily be the case. Unfortunately, different years and companies have chosen slightly different *html* structure of the documents, as well as naming, ordering and etc. What's more, some reports simply do not have any MD&A section at all (confirmed by manual checking). The regular expression for the extraction provided in this paper may not have the highest recall, and that is what could be considered as one of the potential further improvements of the paper. The extraction managed to reach 85% accuracy, which translates into over 230 000 successfully detected MD&A sections out of 276.000 documents (keep in mind that not all of them actually have MD&A, but it's complicated to estimate this share). Thus 85% accuracy is an estimation from below.

Once all the MD&A sections are extracted, the work with text begins. The steps include the following:

1) Dropping tables from text. Be it the MD&A section or any other part of the report, tables contain numbers and are enclosed in *html* markdown tags and are of no use for sentiment evaluation
2) Text is:
    a) Lowercased
    b) Unicode symbols are removed
    c) All punctuation symbols are removed
    d) Newline/tab special symbols are dropped
    e) Stop-words are got rid of
    f) The words are stemmed according to PorterStemmer [30]

One may argue that lemmatization could be a better choice, but stemming has the speed advantage, which is valuable dealing with such large data volumes.

B. Descriptive Analysis

Before moving to modeling, it is always a good idea to get more familiar with the underlying data. Firstly, consider how the number of reports varies depending on year, along with mean absolute return. This gives an idea that more and more companies are being registered on stock exchanges within the years with the relation looking linear. Secondly, In terms of mean absolute return it is easy to cross validate some notorious events. Undoubtedly, subprime mortgage crisis of 2008 resulted in high market volatility, which we are able to see on the chart clearly. Consequences of dot-com bubble in early 2000s are visible as well. Last but not least, COVID-19 hit the world hard in 2020, causing the highest (within the considered range) mean absolute returns among the US traded firms.

Top-2 sector by the number of published reports are Healthcare and Financial Services. Technology only stands on the 3rd place with similar number of reports with Industrials, Consumer Cyclical. There's a more than 50% cut by the reports count down to the 6-th place in the ranking, taken by Real Estate. The least popular sector being Communication Services with less than a thousand filings over 20-year period in the sample. Note that there's also a small share of reports with no sector specified, but that can be neglected.

Cross validating the results with notorious crisis's in the past makes total sense. After continuous decline of 2000-2002 in Technology sector due to the Dot-com bubble, 2003 features rapid growth with mean return exceeding 20%. Rapid growth in that year also occurred in the Healthcare sector, but we can also see that the market did well to recover, reaching 10-15% average quarterly return consistently across all sectors. White stripe of 2008 signals that no sector managed to come out of mortgage crisis unhurt. Finally, there is a clear evidence that COVID-19 had a profound impact of development of Technology and Healthcare sectors. Recall the success of Zoom and all the companies who managed to invent the vaccines during that year.

C. Sentiment Features

The very heart of the current paper is of course sentiment features. This chapter elaborates on their computation logic. Despite several papers claiming Harvard GI dictionary to be less useful in financial application than the one developed by [22], author decides to still use it. The rationale for such an unpopular choice is the desire to go further than just computing the "net sentiment", being the difference between Positive and Negative words driven scores, but to also consider usage of other emotional categories. Overall, Harvard GI offers to split

words into over 100 categories, while for the sake of this research author picks 13 of them. Note that some categories have an intersection in terms of words they contain, which makes total sense. Indeed, vice words ought to be a subset of a broader segment – negative.

In comparison to the paper by [25], author does not just count the frequencies of positive and negative words to compute the net sentiment index afterwards. Instead, words are assumed to carry different importance. Even from the common sense, words that occur in multiple reports frequently (such as stop-words, although they are removed int the paper), should carry less importance to explain the emotional context of a particular text piece. There is a very well-known framework to model compute weights for words within a set of documents: *TF-IDF*. Below author presents a little theory behind the approach. *TF-IDF* consists of two main components:

- Term Frequency (*TF*). It is responsible for counting frequency of each word occurrence in each of the documents, which are represented as MD&A parts in the paper. More frequently encountered words in an MD&A receive higher weight.

$$TF_{i,j} = \frac{n_{i,j}}{\sum_{j=1}^{k} n_{i,j}} \tag{1}$$

  $n_{i,j}$ represents the number of times that term i occurs in document j and k is a total number of documents in dataset.
- Inverse Document Frequency (*IDF*). This part if responsible for lowering weights for words which occur in large number of documents. Indeed, this lowers the ability of a word to convey document-specific message.

$$IDF(d_{i,j}) = log\left(\frac{N}{df_i}\right) \tag{2}$$

  where $d_(i,j)$ stands for $i-th$ word in $j-th$ document; df_i - number of documents containing $i-th$ word; N - total number of documents.
- Finally,

$$TF-IDF_{i,j} = TF_{i,j} * IDF_{i,j} = \frac{n_{i,j}}{\sum_{j=1}^{k} n_{i,j}} * log\left(\frac{N}{df_i}\right) \tag{3}$$

The implementation of *TF-IDF* is not straightforward due to having a huge dataset. None of out-of-the-box solutions suited, so author implemented the following logic in *PySpark* [31]:

1) Preliminary dataset
   - compute word frequencies for each MD&A
   - explode the word-frequency pairs
   - save/persist the file
2) Read the doc-word-frequency table and compute *IDF* (given that now for each word it is known which documents it appears in) via window functions (joins are not advised in order to omit shuffling the data excessively)

D. Main Dataset – Other Features

Once *TF-IDF* scores are computed for each word-document pair, stems of Harvard GI dictionary are joined to assign corresponding emotional categories (of course, multiple categories per word are allowed). Only then the rest dataframes are joined:

- metadata dataframe with information regarding sector, country, industry, delisted Boolean etc.
- stock prices data

The features extracted from stock prices are basically the previous quarters' performance in between the report publication date and 1 day prior to the subsequent quarter's filing post. The same logic applies to the target. There is a trade-off between how many lags to include in order to potentially get a better model and the volume of data. Experiments showed that there is no significant benefit of including more than 4 lags, so that is the final call. Needless to say, all companies with no consequent periods of 4 quarters data are removed.

IV. Methods

Building a regression model for stock prediction is a complicated task. Especially taking into consideration the prediction horizon of a quarter. No doubts that this is a long time for various events to occur, which could outweigh the report sentiment driven momentum drastically. The results would also depend on outlier preprocessing heavily, as well as on the learning algorithm's loss function. In order to have a more robust conclusion regarding the extracted sentiment usefulness, author decided to convert the prediction task to a binary classification. The period P_{target} for the target is between report publication and one day prior to the subsequent quarter's filing post. In the target value 1 corresponds to situation when return during P_{target} is positive and target value 0 otherwise.

Next, a crucial step is determining the logic of splitting the dataset in training and test samples. Potentially, there may be two approaches. First is splitting by companies (i.e.: selecting a set of companies for training and validating the results on the complement). Another approach is to proceed similar to [24], who splits the dataset by time (i.e.: training on data up to some date and validating on more recent observations). The first approach clearly has a danger of having a peeking problem. For instance, in case there appears a confounding factor causing some feature to be inflated within a period, the model could catch it and result in better accuracy on the training set, although the true generalization would be harmed. Additionally, what author considers is more interesting to see is not the ability to extrapolate knowledge of sentiment impact across firms, while to see if the conclusions hold in time. Therefore, training dataset is taken so that it occupies approximately 85% of the least recent data points. The remaining is taken to be the test set. *Catboost* [32] algorithm is first trained on lagged return features only. Consequently, the performance which is represented by AUC is compared with a model with sentiment features included. Note that because there is no severe imbalance in the dataset, AUC is a valid choice for evaluating model performance.

TABLE I
MODELS COMPARISON BY USED FEATURES (INCLUDED FEATURES ARE LISTED)

Model Features	**AUC**	**AUC 95% confidence interval**
Return Lags	0.5324	[0.5182, 0.543]
Return Lags + Sentiment	0.5523 (+3.7%)	[0.5442, 0.5609]
Return Lags + Sentiment + month	0.5544 (+4.1%)	[0.5454, 0.5612]
Return Lags + Sentiment + month + sector	0.5627 (+5.7%)	[0.5536, 0.5696]

TABLE II
BASELINE CLASSIFICATION REPORT

	precision	recall	fl-score	support
False	0.58	0.28	0.37	13767
True	0.51	0.79	0.62	13077
accuracy			0.53	26844

TABLE III
SENTIMENT AUGMENTED MODEL CLASSIFICATION REPORT

	precision	recall	fl-score	support
False	0.62	0.3	0.4	13767
True	0.52	0.81	0.63	13077
accuracy			0.55	26844

V. RESULTS

Top-level results of model fits are presented in the Table I. Although the model quality metrics do not look outstanding by the absolute numbers, they are nevertheless meaningful. First of all, note that the testing set is large enough to assert the quality is significantly better than random guessing. Secondly, model augmented with sentiment outperforms the one on returns lags solely. Due to 95% confidence intervals having no intersection, author concludes that sentiment indeed carries useful incremental information to improve next quarter stock predictions. Curiously, built-in *Catboost* feature importance plot shows that *Weak* emotional category tops the ranking, outweighing the importances of all previous quarters' returns. Once the main hypotheses has been validated, there are several more features ready to add to the model from the current dataset. For instance, sector and month of the report publication. Potentially, returns closer to year end could be higher ceteris paribus, challenging market efficiency. In fact, [33] conducted a study to prove that returns preceding holidays are unusually high (which actually is proven to hold regardless of time of the year). Sector was actually one of the variables [24] suggested adding as a potential improvement. It can be seen that adding *month* variable does not help a lot, yielding just about 0.4 percentage points over the sentiment augmented model. Sector, in contrast, is seen to drive a notable improvement with AUC exceeding 0.56.

In more details, see the baseline classification report (Table II). All the classification reports are presented for the threshold value of 0.5. Adding sentiment features to the baseline results in modest though significant improvement in the primary metric of AUC. Reader can see that both classes show a positive change in *Precision* & *Recall* simultaneously (Table III).

Enriching the model with *Month* feature has minor effect on *AUC*, as well as on more granular metrics. There is a trade-off between Positive & Negative classes' *Recall*, while *Precision* remains the same.

Finally, *sector* feature makes both *Precision* & *Recall* climb up just a little bit. Combined with *month* feature, the latest model outperforms the III by the means of higher *Recall* on Negative class.

VI. CONCLUSION AND FUTURE WORK

The paper focuses on the incremental effect of sentiment emotions extracted from regular 10-K and 10-Q reports' MD&A section on predicting subsequent quarter stock price returns. All US-based companies traded within the last 20 years are parsed, and MD&A sections are extracted using efficient and accurate regular expressions. *TF-IDF* technique is employed to assign weights to words, considering only word stems matching selected emotional categories from the Harvard IV-4 General Inquirer. The overall scores for each category are obtained by summing their respective weights. A *CatBoost* model is fit on data with and without sentiment features to check if the AUC metric experiences any changes. Regression models are avoided due to heavy preprocessing reliance, and binary classification provides more robust results. The model outperforms random guessing significantly, with MD&A emotions contributing to significant improvement in quarterly return lags. Positive and negative sentiment scores show little correlation and importance; instead, more precise sentiment categories, such as *Weak* & *Underst*, prove to be more useful in terms of feature importance.

Future papers can consider improvements, such as incorporating accounting fundamentals into the model, which may reduce the importance of text-driven features as MD&A may contain related sentences. Additionally, the dictionary used in this research is known for financial applications' drawbacks [22], but no financially-sound dictionary with word categories other than positive/negative exists. Further investigation into specific emotions' impact remains the core focus.

From a data perspective, there are ways to enhance the work. Some quarterly reports were missing, possibly due to data issues on the SEC side or other factors, indicating that the filings do not exist. MD&A extraction could be more thorough, considering that some filings lack MD&A or have it under various keywords and section numbers. Manual rule development or fitting a model for automatic extraction could be explored, though collecting a dataset for the model poses a challenge. Moreover, not every filing is available in *html* extension, making parsing text a more complicated task.

ACKNOWLEDGMENT

The work of I. Makarov on Section II was supported by the RSF under grant 22-11-00323 and performed at HSE University, Moscow, Russia in 2023. The work was revised in lie with the conference feedback in 2024.

REFERENCES

[1] H. Z. Kofarbai and M. Zubairu, "Efficient market hypothesis in emerging market-a conceptual analysis," *European Scientific Journal, ESJ*, vol. 12, no. 25, p. 260, Sep. 2016.

[2] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[3] H. Lee, M. Surdeanu, B. MacCartney, and D. Jurafsky, "On the importance of text analysis for stock price prediction," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 1170–1175.

[4] "General inquirer," https://inquirer.sites.fas.harvard.edu/, 2023, online.

[5] R. Laptsuev, M. Ananyeva, D. Meinster, I. Karpov, I. Makarov, and L. E. Zhukov, "Information propagation strategies in online social networks," in *International Conference on Network Analysis (NET'17)*, ser. PROMS, National Research University Higher School of Economics. Berlin, Germany: Springer, June 22–24 2017, pp. 319–328. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-96247-4_24

[6] A. Zaynutdinova, D. Pisarevskaya, M. Zubov, and I. Makarov, "Deception detection in online media," in *Proceedings of the 5th Workshop on Experimental Economics and Machine Learning (EEML'19)*, National Research University Higher School of Economics. Cham, Switzerland: CEUR Workshop Proceedings, September 25 2019, pp. 121–127. [Online]. Available: http://ceur-ws.org/Vol-2479/project5.pdf

[7] R. M. Khayrullin, I. Makarov, and L. E. Zhukov, "Predicting psychology attributes of a social network user," in *Proceedings of the 4th Workshop on Experimental Economics and Machine Learning (EEML'17)*, TU Dresden. Cham, Switzerland: CEUR Workshop Proceedings, September 17–18 2017, pp. 2–8. [Online]. Available: http://ceur-ws.org/Vol-1968/paper1.pdf

[8] L. Sherstyuk and I. Makarov, "Context-based text-graph embeddings in word-sense induction tasks," in *Recent Trends in Analysis of Images, Social Networks and Texts*. Cham: Springer International Publishing, 2022, pp. 68–81.

[9] K. Tikhomirova and I. Makarov, "Community detection based on the nodes role in a network: The telegram platform case," in *Proceedings of the 9th International Conference on Analysis of Images, Social Networks and Texts (AIST'20)*, ser. LNCS, Skoltech. Berlin, Germany: Springer, October 15–16 2020, pp. 294–302.

[10] A. Pugachev, A. Voronov, and I. Makarov, "Prediction of news popularity via keywords extraction and trends tracking," in *Proceedings of the 9th International Conference on Analysis of Images, Social Networks and Texts (AIST'20)*, ser. CCIS, Skoltech. Berlin, Germany: Springer, October 15–16 2020, pp. 37–51.

[11] D. Chun, H. Cho, and D. Ryu, "Economic indicators and stock market volatility in an emerging economy," *Economic Systems*, vol. 44, no. 2, p. 100788, 2020.

[12] B. S. Paye, "Predictive regressions for aggregate stock market volatility using macroeconomic variables," *Journal of Financial Economics*, vol. 106, no. 3, pp. 527–546, 2012.

[13] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grč ar, and I. Mozetič, "The effects of twitter sentiment on stock price returns," *PLOS ONE*, vol. 10, no. 9, p. e0138441, sep 2015.

[14] L. Bing, K. C. Chan, and C. Ou, "Public sentiment analysis in twitter data for prediction of a company's stock price movements," in *2014 IEEE 11th International Conference on e-Business Engineering*, 2014, pp. 232–239.

[15] R. Feldman, S. Govindaraj, J. Livnat, and B. Segal, "Management's tone change, post earnings announcement drift and accruals," *Accounting*, 2009.

[16] A. K. Davis and I. T. Sweet, "Managers use of language across alternative disclosure outlets: Earnings press releases versus md&a," *Contemporary Accounting Research*, vol. 29, no. 3, pp. 804–837, 2012.

[17] A. K. Davis, J. M. Piger, and L. M. Sedor, "Beyond the numbers: Measuring the information content of earnings press release language," *Contemporary Accounting Research*, vol. 29, no. 3, pp. 845–868, 2012.

[18] F. Li, "Annual report readability, current earnings, and earnings persistence," *Journal of Accounting and Economics*, vol. 45, no. 2, pp. 221–247, 2008, economic Consequences of Alternative Accounting Standards and Regulation.

[19] S. B. Bonsall, A. J. Leone, B. P. Miller, and K. Rennekamp, "A plain english measure of financial reporting readability," *Journal of Accounting and Economics*, vol. 63, no. 2, pp. 329–357, 2017.

[20] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *Journal of Finance*, vol. 23, pp. 589–609, 1968.

[21] G. Andreeva, J. Ansell, and J. Crook, "Credit scoring in the context of european integration: is there a future for generic models?" *Journal of Financial Transformation*, vol. 23, pp. 129–134, 2008.

[22] T. Loughran and B. McDonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-ks," *Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011.

[23] P. Gandhi, T. Loughran, and B. Mcdonald, "Using annual report sentiment as a proxy for financial distress in u.s. banks," *Journal of Behavioral Finance*, vol. 20, pp. 424 – 436, 2018.

[24] W. Lee and B. Suh, "Modeling stock prices with text contents in 10-q reports," in *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 2018, pp. 224–229.

[25] J.-H. Meier, W. Esmatyar, and R. Frost, "The predictive power of the sentiment of financial reports," in *ICTERI Workshops*, 2018.

[26] D. Yermack, "Good timing: Ceo stock option awards and company news announcements," *The Journal of Finance*, vol. 52, no. 2, pp. 449–476, 1997.

[27] D. E. Allen, M. McAleer, and A. K. Singh, "Daily market news sentiment and stock prices," *Applied Economics*, vol. 51, no. 30, pp. 3212–3235, 2019.

[28] "Sharadar dataset," http://www.sharadar.com, 2023, online.

[29] R. Aroussi, "yfinance Python library," 2023, [Online].

[30] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 40, pp. 211–218, 1997.

[31] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin *et al.*, "Apache spark: a unified engine for big data processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016.

[32] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 6639–6649.

[33] G. N. Pettengill, "Holiday closings and security returns," *Journal of Financial Research*, vol. 12, pp. 57–67, 1989.