Using a Natural Language Processing Approach to Predict Future Stock Market Performance Based on Existing Twitter Sentiment.

Maxwell Klaiman1,2

1. Northwestern University School of Professional Studies

Master’s of Science in Data Science Program

1. Church St, Evanston, IL 60201

2. Allina Health System

2925 Chicago Avenue, Minneapolis, MN 55407

Address to which correspondence should be addressed:

[MaxwellKlaiman2023@u.northwestern.edu](mailto:MaxwellKlaiman2023@u.northwestern.edu)

**Abstract**

The relationship between consumer sentiment and stock index performance is often touted. However, the nature of that relationship is poorly established in standardized investigation and prevailing literature. Natural Language Processing, NLP, can be used for a wide variety of opinion-gathering methods, including sentiment analysis. Perhaps intuitively, sentiment analysis seeks to quantify the prevailing opinion from a provided sentence, and can be used to classify text as positive, negative, or neutral. The asserted relationship between the stock market – notably the Standard and Poor’s 500 index - and the sentiment of the public around those companies can be investigated using deep learning techniques. A 3x1 design of experiments, looking at three NLP packages and two neural network classifiers, seeks to determine which existing tools most accurately reflect the relationship between consumer sentiment and stock price for the purpose of future prediction. Success in quantifying this relationship could be crucial, and has implications for shareholders, businesses, and public interest as a whole.

**Keywords:** Natural Language Processing, Deep Learning, Sentiment Analysis, Neural Network, S&P 500, Artificial Intelligence, Stock Market

1. **Introduction**

The Standard and Poor’s 500 (commonly referred to as the S&P 500) is a key indicator of the health and current value of high-value American stocks. Comprised of a weighted average of the 500 largest publicly traded American companies, the index reflects the expansion or contraction of market values as a general trend, as well as the isolated performance of individual stocks. Predicting future stock market prices is an area of key interest to both individual investors and corporations – largely due to the ability of effective predictions in making money for investors, companies, and their shareholders (Puh 2023). The value of an effective market prediction lies in both the ability for investors to earn, but arguably more importantly the potential for organizations to avoid non-trivial losses. In 2022, for example, American’s holdings fell by $9 Trillion – a loss with catastrophic and continuing consequences to the American workforce (Frank 2022). While these losses may have been unavoidable, well-informed investment practices may have allowed investors and corporations to avoid hardship or layoffs.

Perhaps intuitively, there is a relationship between stock pricing and consumer sentiment, which has been stated repeatedly but is poorly investigated. According to Liu, et al. (2023), repeated economics studies have been conducted to look at a variety of sentiment-driven effects, including the “day of the week effect,” along with other more general sentiment around liquidity, values, and crises. In recent work, much attention has been given to the sentiment analysis conducted, indicating the possible predictive value of recent Natural Language Processing (NLP) techniques.

1. **Literature Review**

Attempts at utilizing sentiment analysis techniques to project stock market performance have been relatively common, largely due to the financial implications of an effective model (Liu 2023). Casting a wide net over recently-published articles, a wide variety of methods to conduct sentiment analysis are utilized in attempts to categorize the relationship between sentiment and market performance. Wonseong Kim (2023) uses Word2Vec categorization and a bi-directional long-short term memory (BILSTM) neural network to connect the conducted sentiment analysis to market returns. Kim’s research states that an enhanced lexicon correlates with improved sentiment analysis accuracy, but there is no direct statement of model accuracy.

Patel, et al. (2021) performed a more advanced analysis of the different models that can be used to predict future stock prices, including LSTM, XGBoost, ADABoost, and a basic Random Forest classifier. LSTM substantially outperformed each of the other algorithm types. However, even the LSTM model was a very poor predictor of future stock market (Dow Jones) values. It is noteworthy that little attention in these cases was given to how the sentiment analysis itself was conducted.

There exist many options for conducting natural language processing on a dataset, which are well-categorized by Birjali, Kasri, and Beni-Hassane (2021). Some key NLP package options are TextBlob, NTLK, CoreNLP, and OpenNLP (via NTLK), of which TextBlob appears to be the highest performing. The authors make special note of the possibility to leverage Convolutional Neural Network (CNN) practices to perform sentence-level sentiment classification.

Bing Liu, in his 2012 book *Sentiment Analysis and Opinion Mining*, goes into some of the challenges faced in conducting sentiment analysis, most notably in detecting sarcasm and mixed-sentiment words. In his example, Liu mentions the word “sucks” as a possible mixed-sentiment word, where “This camera sucks” and “This vacuum cleaner really sucks” have opposite sentiments. To counteract this, sentence-level (or greater) classification must take place when conducting sentiment analysis.

To address the issues with word-level classification, two additional techniques come into play: VADER and FinBERT. VADER is a rule-based text analysis technique, which is able to assign a compound sentiment value to text (Elgabib and Ying 2019). Additionally, VADER is able to produce high-accuracy analysis with even a small training sample, which lends itself well to a smaller sample. Of perhaps even greater interest is FinBERT, a financial text analysis model based on BERT (Bidirectional Encoder Representations from Transformers). Transformers, at their core, are able to leverage bidirectionality which provides an advantage when compared with traditional NLP algorithms (Araci 2019). FinBERT is able to leverage bidirectionality, as well as its specific financial language training, to provide the highest-fidelity output (Huang 2022). While there are other specialized BERT-based models, the relevance of FinBERT to financial sentiment analysis makes it a prime candidate for conducting this analysis.

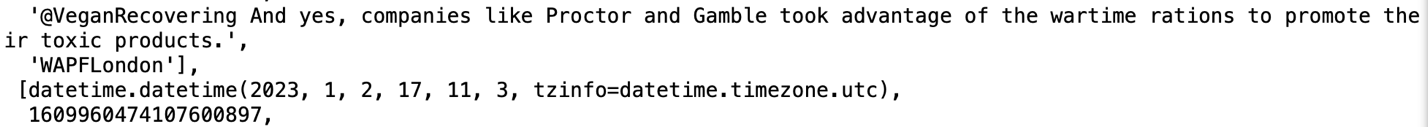
1. **Data**

Recent developments in Twitter’s policy has made their access point interface (API) nearly un-usable without expensive enterprise licensing. As such, Twitter data will have to be scraped using a web crawler. The data to be analyzed will be obtained using the TwitterSearchScraper Python module, pertaining to a selection of relevant S&P 500 contributors. The 14 highest-weight companies in the S&P 500 are, in order:

* Apple
* Microsoft
* Amazon
* Nvidia
* Alphabet
* Berkshire Hathaway
* Meta
* Exxon Mobil
* UnitedHealth Group
* Tesla
* Johnson & Johnson
* JPMorgan Chase & Co.
* Visa
* Proctor & Gamble

(“S&P 500 Companies by Weight”). As such, no fewer than ten tweets per day pertaining to each of these companies will be scraped via TwitterSearchScraper, and used to establish a corpus of tweets that contain sentiment-rich information about the largest companies in the S&P 500.

Additionally, some metadata around the tweets needs to be collected in order to assess the temporality of information and some basic tweet identification. In short, the entirety of data collected from Twitter is (1) full text of the tweet, (2) tweet date and time, and (3) tweet ID. While additional information can be pulled (i.e. Twitter handle), only directly relevant information will be pulled for the sake of simplicity. An example of the information that can be pulled, prior to processing and cleansing, is below:



In addition to the sentiment analysis, S&P 500 data will need to be pulled and aggregated in a target range, which allows for both reasonable computational analysis and generalizability of the results. These dual objectives are at odds, which is only furthered by limitations in Twitter’s API preventing straightforward, low-cost web scraping. As such, a relatively small window must be established. For the sake of this analysis, a 3-month (one quarter) window will be established, with day-by-day breakdowns of S&P 500 open & close price. This window will run from the start of Q1 2023 to the end of Q1 2023 (January 1 to March 31, 2023). Presumably, this will allow for enough data for effective NN training, without burdening resources.

1. **Methods**

In order to develop the most effective algorithm at predicting future stock market prices based on twitter sentiment analysis, a holistic investigation must be conducted of both sentiment analysis techniques and classifier types. Due to the incredibly broad scope that can be applied to an analysis of this type, some limitations must be set. As such, a specific experimental design, with one main classifier types and three NLP tools will be conducted, and analyzed against S&P 500 performance over a 90-day window from the start to finish of Q1 2023.

The broad type of classifier that will be used in classification is Recurrent Neural Networks – specifically Long Short Term Memory (LSTM). Since bidirectional LSTMs (BiLSTM) have been shown to outperform all other classifiers, only a BiLSTM will be developed, tuned, and used for classification. For the NLP portion of the design, three major sentiment analysis techniques will be utilized in the three by one experimental design. These sentiment analysis techniques are using FinBERT packages, TextBlob, and VADER.

Each of these tools comes equipped with the tools to address one of the most straightforward challenges faced by NLP algorithms: understanding the meaning of a word within its larger context. TextBlob uses sentence-level contextualization to evaluate the polarity and objectivity of a sentence, which can be leveraged into a compound score. VADER, in a similar fashion, leverages a straightforward rule-based model to predict sentiment in an accurate manner. Perhaps most promising, FinBERT is a finance-trained, BERT-derived transformer. It is designed to analyze the sentiment around financial information specifically, and is expected to do so in a manner which outperforms both VADER and TextBlob. Textblob, which is the oldest algorithm being investigated, will serve as a performance baseline of sorts for this task.

Major considerations must be given to the evaluation of each sentiment analysis technique, particularly to ensure that items are not blatantly mislabeled. This presents a conundrum, as minor-to-moderate differences between the sentiment assignment of these tools will form the foundation of a distinction between algorithms. However, major discrepancies between assigned and actual sentiment will be resolved to reflect the actual sentiment of each tweet in question. To determine the minimum threshold for a major discrepancy, tweets that are incorrectly marked with a negative tone will be reverted to positive, and vice-versa. Tweets marked as neutral which are not clearly positive or clearly negative will be left untouched.

From there, the final sentiment generated by each algorithm will be passed into a bidirectional LSTM, which is being trained to predict S&P 500 closing price of a small set of dates following the training dates. In addition to each model being tested for direct comparison, the model which most accurately predicts market prices will be tested in a low-training and high-training environment to ascertain the impact of additional training epochs.

1. **Results**

With the potential for an unmanageably large dataset to be developed, considerations were made with regards to the computational capacity of a laptop computer. In all, around 15,000 tweets were scraped from Twitter, each pertaining to one of the top 14 S&P 500 companies, ranging from January 1, 2023 to March 31, 2023. There were a minimum of 10 and a maximum of 20 tweets each day relating to each of the 14 companies. To assess the predictability of the S&P 500 from this limited corpus. Sentiment analysis was performed using FinBERT, TextBlob, and VADER, respectively. From here, sentiment was condensed into an average score, corresponding to each day the US Stock Market was open. After preparation, S&P 500 Open price and daily aggregate sentiment was passed into a bidirectional long short term memory (LSTM) model for close-price prediction. A typical experiment ran for 3,000 epochs due to computational limitations. To consider the impact of training epochs, the most-effective model (VADER) was trained for both a low and high number of epochs. The results of these predictions are in table 2 below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Performance by Classifier | Number of Epochs | Optimal  Val. Accuracy | Model Loss (Mean Squared Error unless noted) | Run-time |
| FinBERT | 3000 | 0% | Accuracy: 1.82% | 00:01:58.77 |
| TextBlob | 3000 | 8.33% | Accuracy: 41.82% | 00:02:27.10 |
| VADER – short training | 300 | 8.33% | Accuracy: 0.0562 | 00:01:20.98 |
| VADER – typical training | 3000 | 8.33% | 116,240 | 00:13:00.88 |
| VADER – long training | 20000 | 8.33% | 98,264 | 1:22:31.91 |

Image 1a and 1b, below, show (1) the accuracy and loss of the VADER-based model and (2) the predictive capacity of the model after 3000 training epochs.

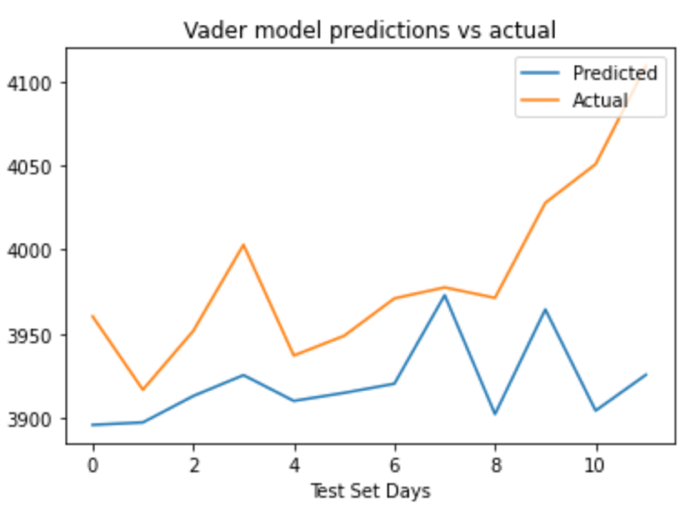
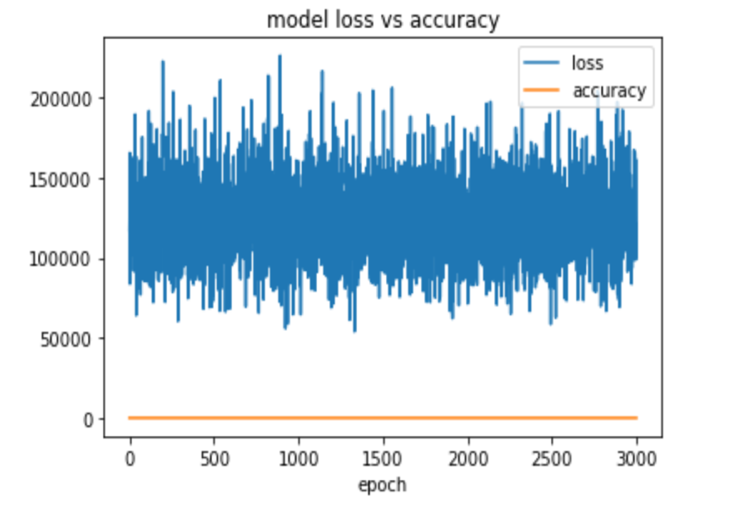
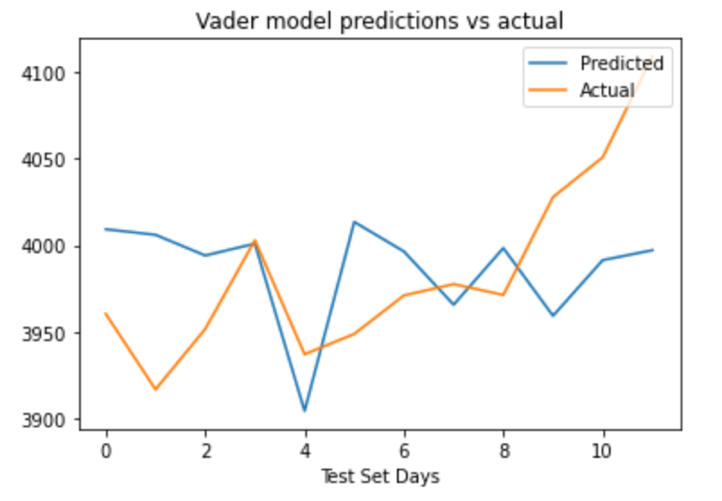
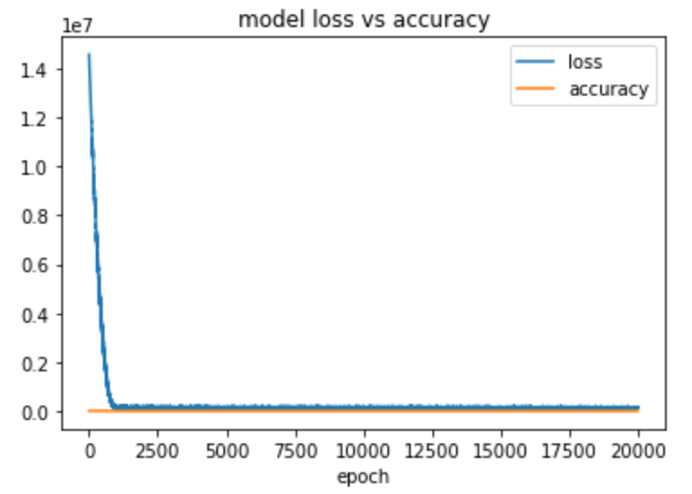


Image 1(a) Image 1(b)

Finally, the results of training the VADER model after 20,000 epochs, the most effective model, are below in Image 2.



1. **Analysis and Interpretation**

In investigation of each of the training algorithms employed to conduct this prediction, a couple of interesting trends become apparent. First off, whether due to the size of the corpus developed or a lack of quantitative data to combine with a sentiment analysis, the predictive capacity of sentiment analysis alone is incomplete if not altogether poor. The highest optimal Validation Accuracy, accomplished by the TextBlob and VADER models, was a mere 8.33% (1 in 12). From Image 1(b), this appears to be mirrored in the test set, where only day 7 is a close match between predicted and actual value. However, this general view of the limited effectiveness of a sentiment-driven prediction algorithm is somewhat reductionist, and a key success can be easily overlooked: the general accuracy of market direction, even in the absence of magnitude. The standard-trained VADER model was able to predict the direction of market change in 9 of the 12 days in the test set – a 75% success rate. This disparity, in which direction is accurately predicted while magnitude is not, speaks slightly to the predictive value of public sentiment in future market performance. Perhaps surprisingly, the highly-trained VADER model did not lock-step direction as closely as the standard training level, which may be indicative of overfitting.

Of perhaps larger importance is the difference between experimental designs and the nuanced capability of FinBERT, TextBlob, and VADER to accomplish a viable sentiment analysis. FinBERT, a transformer-based tool, was expected to perform the most optimally of all analysis methods. However, astonishingly, FinBERT has the lowest relative success in predicting market price based on sentiment alone. This may be due to some slight differences between FinBERT use in this case and its intended purpose. While FinBERT was developed to operate in the Finance world, there are distinct differences between financial reporting in journals and sentiment-driven language on Twitter. One example which illuminates an issue faced by the FinBERT algorithm was its classification of the following sentence, which the algorithm classified as neutral:

"… I love dividend stocks. Been investing into them since 2006. Costco, Proctor & Gamble, Pepsi are some of my biggest holdings."

Upon deeper review, this tweet appears to be one of very few misses by the algorithm, but it seems that the threshold for a non-neutral classification may be higher-than-optimal.

Finally, in looking at the TextBlob Sentiment analysis performance, it appears to be a slightly more effective classifier of financial tweet sentiment than FinBERT, but without the performance enhancements present in the VADER algorithms. This may corroborate the observation that the initial weighting of FinBERT was performed on a dataset too esoteric to handle the casual nature of Twitter. It would appear that TextBlob’s sentence-level analysis technique is quite successful in this context, and may be well-suited for Twitter in general. This is confirmed by the high relative accuracy of market prediction using TextBlob: 42%. While this is not an extremely high number, it greatly outperforms FinBERT in this task.

1. **Conclusions**

There are a couple takeaways that can be discerned from this high-level analysis. First, there is the capacity for further verification of the work, especially as it pertains to the size of the corpus being obtained from Twitter. Due to Twitter API access constraints, primarily related to cost, it was not feasible to conduct a more thorough, years-long analysis of an exhaustive corpus. From the relatively low-power perspective established here, it is clear that sentiment alone is not a sufficient predictor of stock market returns. However, that restriction fails to consider ways in which a more comprehensive approach may seek to improve the performance of a sentiment-based predictive algorithm. For example, if the number of tweets was increased substantially (from 15,000 to 1.5 million) and the time window expanded to match, there is a much larger chance that any analysis would outperform those conducted in VADER, TextBlob and FinBERT.

With an increased focus on corpus size as a key predictor of algorithm success, there is an interesting tension between seeking to produce competitive results with a more limited capacity. Building an exhaustive corpus requires time, computing power, and, ultimately, vast sums of money. While there certainly appears to be a race to the top in terms of building the largest possible datasets for analysis, additional consideration should be given to lower-cost but similarly high-performing algorithms when developing novel models.

1. **Directions for future work**

Looking toward future enhancements of this work, a couple advancements can be made on what has already been done. First, a broader meta-analysis of relevant analysis algorithms could prove to be a fruitful exploration of the major tools that are in existence. With the rapid proliferation of large language models and pursuits into artificial general intelligence, the importance of establishing the solid NLP algorithms in a quickly evolving environment would be of great use. Additionally, with the vast number of transformer-driven algorithms (such as FinBERT), a cataloging effort of those algorithms and their respective utilities in Twitter could be of very high value in directing future Twitter-based analyses.

With regards to utilizing sentiment as a sole predictor of stock market performance, there are a couple possible ways to achieve more accurate results that deserve attention. The first is to consider building a more comprehensive corpus, which could involve (1) more companies, (2) the inclusion of financial keywords, such as consumer price index, which influences markets, and (3) a broader timeframe for data aggregation and analysis. With more computing power, or enterprise access to the Twitter API, each of these objectives may be within reach. Finally, it would be worthwhile to undergo a combined effort at predicting market performance, where a quantitative representation of consumer sentiment is only one aspect of a predictive algorithm. In this case, it may be possible that the inclusion of consumer sentiment allows for improved performance versus more traditional market prediction algorithms. In all, development of a larger dataset – whether by inclusion of more Tweets or metadata – would be a worthwhile undertaking for future analyses.

1. **References**

Araci, Dogu. ‘FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models’. ArXiv [Cs.CL], 2019. arXiv. <http://arxiv.org/abs/1908.10063>.

Birjali, Marouane, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. “A Comprehensive Survey on Sentiment Analysis: Approaches, Challenges and Trends.” *Knowledge-Based Systems* 226 (August): 107134. <https://doi.org/10.1016/j.knosys.2021.107134>.

Dang, Nhan Cach, María N. Moreno-García, and Fernando De la Prieta. 2020. “Sentiment Analysis Based on Deep Learning: A Comparative Study.” *Electronics* 9 (3): 483. <https://doi.org/10.3390/electronics9030483>.

Elbagir, Shibab, and Jing Ying. 2019. Review of Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment. In Proceedings of the International MultiConference of Engineers and Computer Scientists . Vol. 1. Proceedings of the International MultiConference of Engineers and Computer Scientists 2019.

‌Frank, Robert. 2022. “Stock Market Losses Wipe out $9 Trillion from Americans’ Wealth.” CNBC. September 27, 2022. <https://www.cnbc.com/2022/09/27/stock-market-losses-wipe-out-9-trillion-from-americans-wealth-.html>.

Huang, Allen H., Hui Wang, and Yi Yang. "FinBERT: A Large Language Model for Extracting Information from Financial Text." Contemporary Accounting Research (2022).

Kim, Wonseong. 2023. “Words That Wound: The Impact of Biased Language on News Sentiment and Stock Market Index.” *ArXiv (Cornell University)*, April. <https://doi.org/10.48550/arxiv.2304.00468>.

Khurana, Diksha, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2022. “Natural Language Processing: State of the Art, Current Trends and Challenges.” *Multimedia Tools and Applications* 82 (July). <https://doi.org/10.1007/s11042-022-13428-4>.

Liu, Bing. 2012. *Sentiment Analysis and Opinion Mining*. San Rafael: Morgan And Claypool.

Liu, Qing, Woon-Seek Lee, Minghao Huang, and Qingjun Wu. 2022. “Synergy between Stock Prices and Investor Sentiment in Social Media.” *Borsa Istanbul Review*, September. <https://doi.org/10.1016/j.bir.2022.09.006>.

Patel, Ramkrishna, Vikas Choudhary, Deepika Saxena, and Ashutosh Kumar Singh. 2021. “LSTM and NLP Based Forecasting Model for Stock Market Analysis.” *2021 First International Conference on Advances in Computing and Future Communication Technologies (ICACFCT)*, December. <https://doi.org/10.1109/icacfct53978.2021.9837384>.

Puh, Karlo, and Marina Bagić Babac. 2023. “Predicting Stock Market Using Natural Language Processing.” *American Journal of Business*, April. <https://doi.org/10.1108/ajb-08-2022-0124>.

Socher, Richard, Yoshua Bengio, and Christopher Manning. 2012. Review of *Deep Learning for NLP (without Magic)*. *Abstracts of ACL 2012* 1 (1).

“S&P 500 Companies - S&P 500 Index Components by Market Cap.” 2022. Www.slickcharts.com. 2022. https://www.slickcharts.com/sp500.