# Classifying Investor Sentiment

## Introduction

Investor sentiment is often a reliable indicator when assessing future market trends and performance. Not only is this sentiment based on years of experience in monitoring and investing in the stock market, but also determines the investor's actions, which can have an effect on the performance of the stock market. If the general sentiment is positive, more investors are likely to invest larger amounts of money, which can build momentum in the stock market during the short term. On the other hand, if the general sentiment is negative, less investors are going to invest and perhaps even liquidate some of their assets, which could result in the market slowing down in the short term. Keeping track of investor sentiment can be very helpful for beginner investors who want to stay up to date with market performance and make informed decisions about future performance.

At a certain point, an investor should attempt to make their own decisions about market performance, and using their own sentiment. To do so, it would be beneficial to understand what factors affect other investors' sentiment and how. In this paper, I will investigate certain market factors which I believe can influence investor sentiment and market performance. By doing so, I hope to identify a method to classify bullish and bearish investor sentiments based on these factors. There are some factors that I will not be able to account for, such as current events and business news, which definitely have an impact on investor sentiment but are very hard to quantify. I will have to keep this in mind as I conduct my analyses.

## Data Acquisition

The data for my analysis will be obtained from multiple sources, and will be in different formats. The investor sentiment data will be acquired from an American Association of Individual Investors (AAII) dataset. I will be pulling this data from Quandl straight into the python script using the Quandl API. The dataset consists of the percentage of investors that are Bullish, Bearish and Neutral towards the stock market for the next 6 months following when the sentiment was recorded. Members of the AAII are polled on a weekly basis and allowed only one vote per member. The dataset also includes S&P 500 data such as the weekly high, weekly low and weekly close. The main variables I am concerned with are the Bull-Bear Spread and the S&P 500 data. The Bull-Bear Spread is the measure of how much the Bullish exceeds the Bearish (% Bullish - % Bearish). Therefore, the higher the Bull-Bear Spread is, the more Bullish

Investors are than Bearish. This will be my class, or dependent variable. I am not concerned with the Neutral values because this is not as helpful as the Bullish and Bearish values when forming an opinion about the market.

Some of the features for my data will be derived from the S&P 500 data from the AAII data set. I will be using the weekly change of the S&P 500 closing price as one of my features, as well as a measure of volatility based on the week's highs and lows. Additionally, I will include features derived from US Dollar Index futures (USDX) and US Treasury Bond futures (USTB). The data for both these variables will be obtained from Wiki Continuous Futures datasets on Quandl. I will be using the weekly change of these two features in my analysis. Lastly, I will include the moving average convergence divergence (MACD) of the Dow Jones Industrial (DJI), the S&P 500 (SP500) and the NASDAQ Composite (IXIC) as features in my dataset. In order to calculate the MACD, I will also need the simple moving averages (SMA) and exponential moving averages (EMA) for both indices. I will be obtaining the daily historic data for both the afore mentioned indices from Yahoo finance.

All of my data will be a time series format. I plan to only use data from January, 2000 onward because I would like to assess investor sentiment towards the stock market in its most recent state and behavior. I have provided a list of links to the datasets at the end of this paper.

## Data Understanding & Preparation

Before diving into the preparation of the data, it is important to understand what the variables are and why they are relevant for this analysis. All instances in my data will be weekly because that was how the investor sentiment survey was conducted. The class variable, which will be the focus of the classification, is the Bull-Bear Spread. The reason I have chosen this variable over the individual percentages of the Bullish and Bearish sentiments is because it captures the levels of both into a single measurable class. As mentioned earlier, the greater the Bull-Bear Spread is, the more Bullish investors are feeling towards the market, which implies a positive performance in the stock market.

The features that will be used to classify the sentiment will be the SP500 weekly change, SP500 week volatility, SP500 MACD, DJI MACD, IXIC MACD, USDX weekly change and USTB weekly change. All change variables will be measured as a percentage change from the previous week since the values are of different magnitudes, but this will also help account for some inflation over the years which is to be expected when dealing with prices over a long period of time.

The USDX and USTB futures are strong indicators of the performance of the American economy. The USDX is a measure of the value of the US dollar relative to the currencies of their most significant trading partners. Therefore the value of USDX futures, along with USTB futures, are often reflective of the future performance of the American economy. The SP500, DJI and IXIC are the three most followed indices in the US stock market. They are the primary indicators of the current performance of the US stock market. Since the MACD is a momentum indicator, it can be used to indicate the performance of these indices in the near future.

The MACD is calculated by subtracting the 26-day EMA from the 12-day EMA. In other words, if the MACD is positive, this means that its security has been performing better over the past 12 days than it has been over the past 26 days. The same concept applies when the MACD is negative; it has been underperforming in the past 12 days compared to the past 26 days. The greater the MACD is (negative or positive), the more the security has been outperforming or underperforming its longer period. The EMA of a security is a weighted moving average that gives more weighting to the recent prices than the SMA, and is therefore used to calculate the MACD. The calculation of the EMA is as follows:

$$EMA = \left(Price_t \times \frac{2}{P+1}\right) + \left(EMA_y \times \left(1 - \frac{2}{P+1}\right)\right)$$

Where *$Price_t$* is the price today, *$EMA_y$* is yesterday's EMA of the security and *P* is the length of the period (usually measured in trading days). I will be using the 12 and 26 day EMAs because they are the standard periods used to calculate the MACD. The EMA of the first period (or trading day) is simply the SMA of that security for the same time period. Therefore I will also need to calculate the SMA for the indices. I will use the moving average function in excel to calculate the SMAs, and then continue to use excel to calculate the EMAs and the MACD since I find it easier than using python. I will not be including the other features of the IXIC and the DJI as I am the SP500 because I do not want to overpopulate the dataset and suffer from the curse of dimensionality. I believe that the weekly indicators of the SP500 are enough for me to include in my analysis.

Once I have successfully created the MACDs in excel, I will start my data preparation process in python. First I will merge the separate datasets with my class dataset on the date since some all other datasets are daily. Next I will calculate the necessary weekly changes (change between each instance) and drop the original closing prices from the dataset. I will be calculating the volatility of the SP500 as follows:

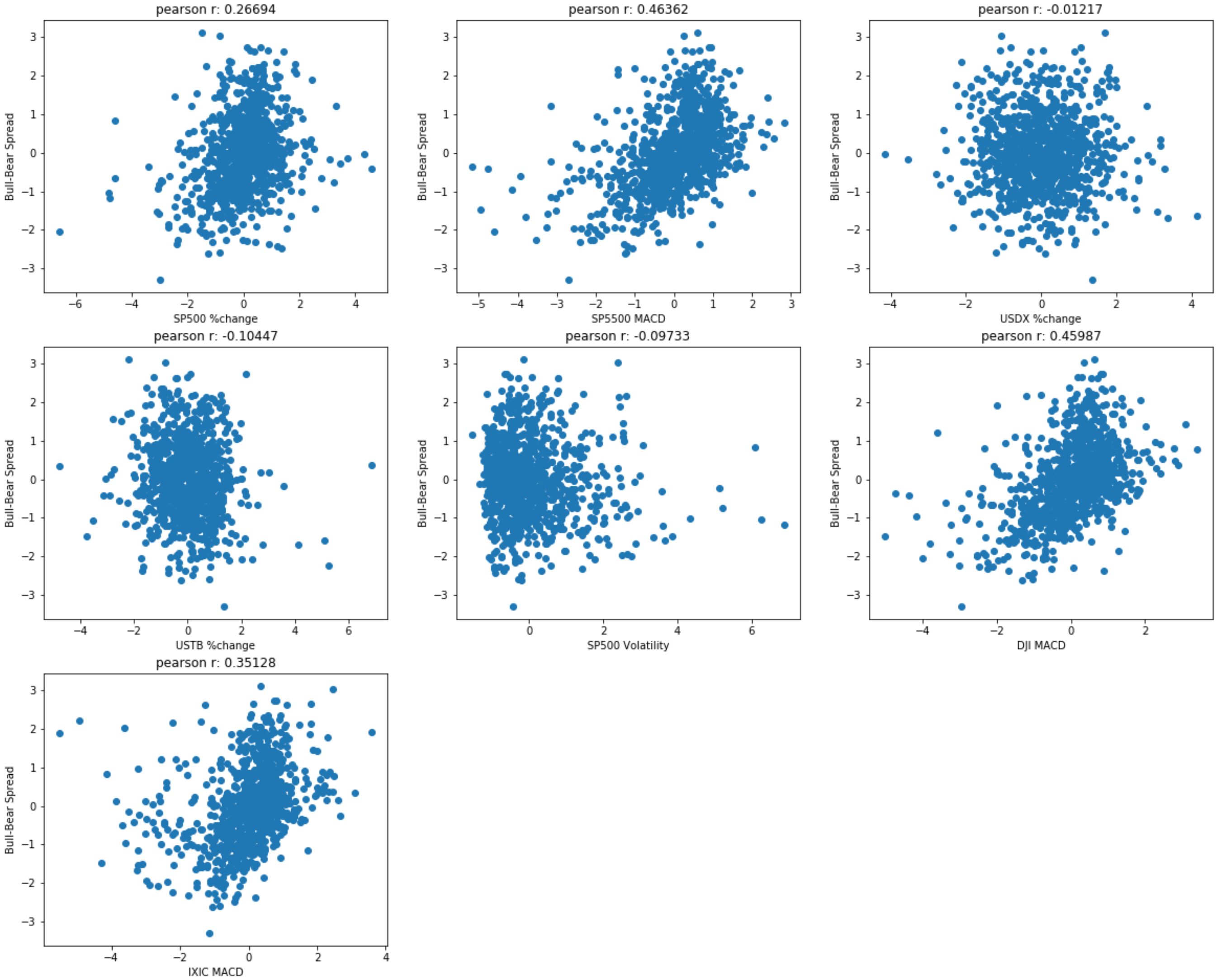
$$v = \frac{High - Low}{Close}$$

The weekly highs and lows are included in AAII data set, along with the weekly close. Although this is not an official method to calculate volatility, I believe it will serve its purpose for my analysis. Once I have done so, I will drop all unnecessary variables, and keep only the features that I have mentioned above. I will then separate the class variable – Bear-Bull Spread – from the features and include it in a separate dataset. Finally, I will scale all the data in order to normalize their range for better visual results.

To gain a better understanding of my data and the relationships between the features and the class, I plotted them individually against the class as is seen in Figure 1. on the next page. It is evident that the MACDs are have the highest correlation to the Bull-Bear Spread. The SP500 MACD is the most correlated (*r* = 0.46362), with the DJI MACD being the next most correlated (*r* = 0.45987) and the IXIC MACD being the third most correlated feature (*r* = 0.35128). The SP500 %change also seems to show some correlation (*r* = 0.26694). The rest of the variables do not seem to exhibit much correlation to the class; however I will still continue to use them in my analysis as they might contribute to some insight when combined with the other factors.

For my final analysis, I will use a combination of clustering and classification methods to further investigate the data. In order to do so, I require my class to be discrete rather than continuous. Therefore, I will convert the Bull-Bear Spread into a discrete variable by separating it into two different bins. These bins will essentially resemble a general bullish or bearish sentiment. The lower bin will resemble instances where the investor sentiment is more bearish, and the upper bin will resemble instances where the investor sentiment is more bullish. This will make the classification of these instances much simpler. I had initially hoped to perform a multiple regression analysis using the features, but I soon realized that my results would be more accurate using a classification method due to the weak bivariate relationships between the features and the class as depicted in Figure 1.

Figure 1:



## Data Analysis: Clustering

The first analysis I will perform on my data will be a clustering analysis. My aim is to identify two clusters in the features that will separate the bullish instances from the bearish instances. I will use the KMeans and Agglomerative clustering algorithms to perform the clustering, and then compare the two models based on a scatter plot visualization. For the agglomerative clustering, I will be using the “complete” linkage in order to capture the furthest possible members of the cluster. For the first clustering analysis, I will include all features when fitting the model. I will plot the results on a scatter plot with the DJI MACD as the y-axis and the IXIC MACD as the x-axis since there are fewer overlapping stocks between the two than there are between SP500 and IXIC, and all of the DJI stocks are included in the SP500. Therefore, by plotting the DJI and the IXIC, I hope to avoid too much similarity in their performances. The results of the first clustering analysis can be seen below in Figures 2. and 3.

Figure 2.

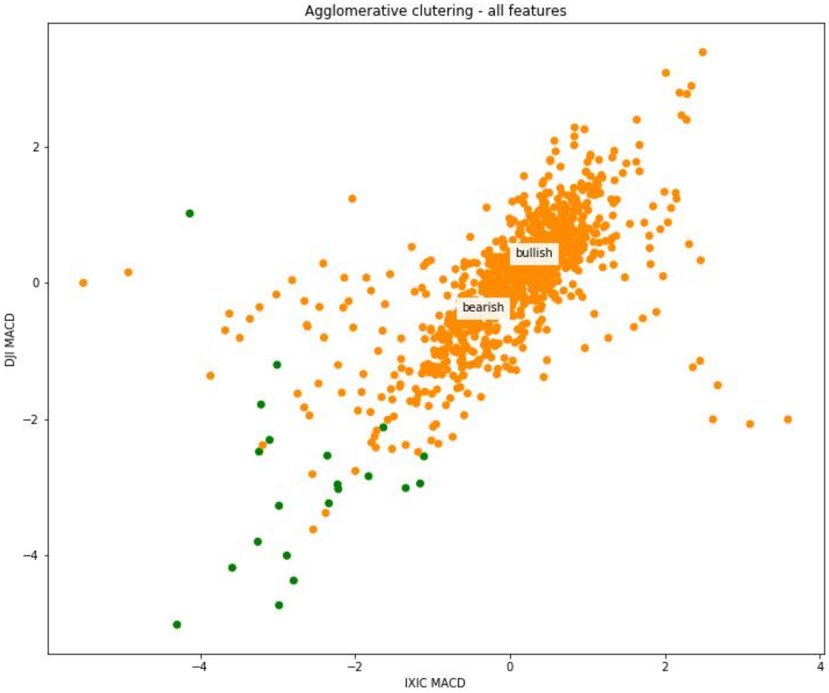


Figure 3.

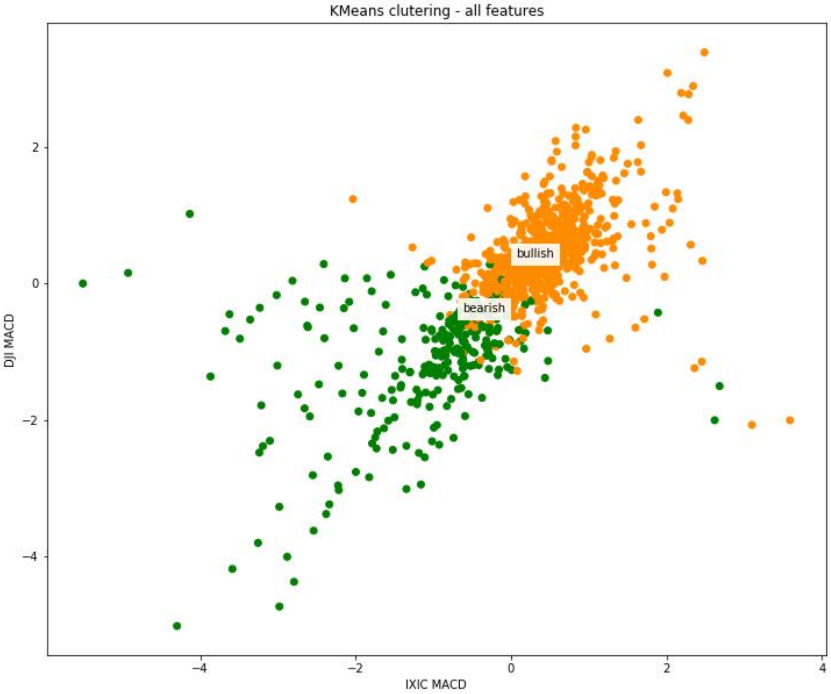
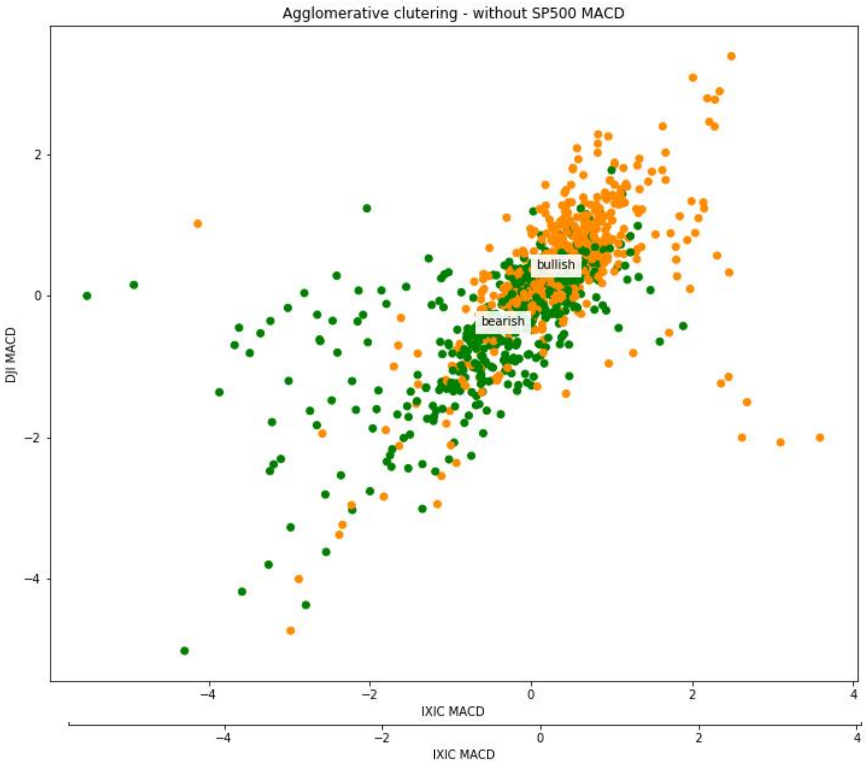


Figure 4. will depict the actual values of the instances

Figure 4.

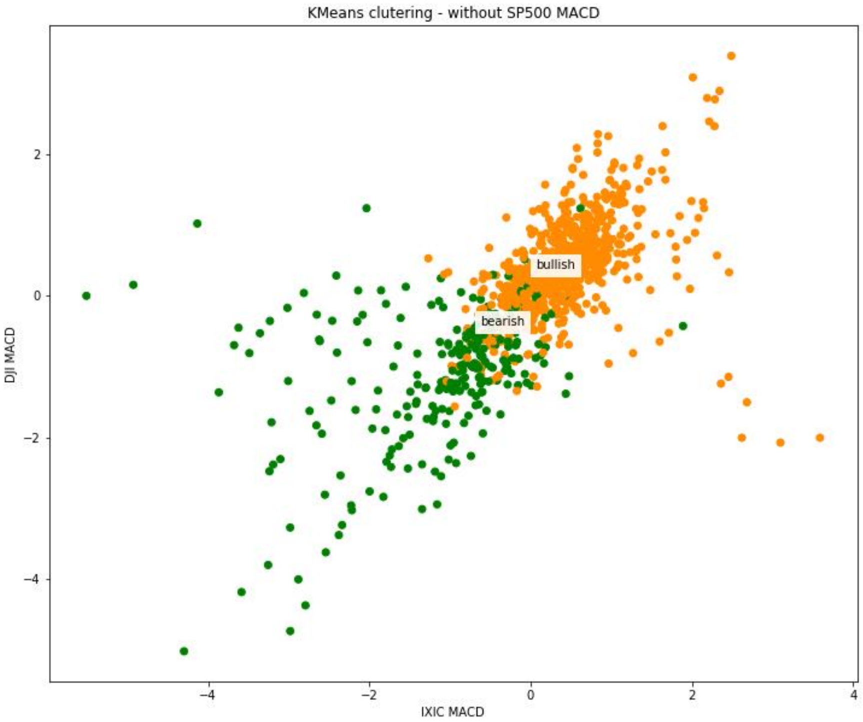


In the visualizations above, the orange instances resemble bullish sentiment, and the green instances resemble bearish values. The means of the actual instances can be seen on the plots in the labelled boxes. By observing the actual values plot, it is evident that there is some separation between the class values: the bullish (orange) instances seem to occur more when the MACDs are above 0, and the bearish (green) occurs more when the MACDs are below 0. Based on the visualizations, it would seem as though the KMeans clustering method more accurately depicts the actual values of the bullish and bearish instances, and is most successful in separating the clusters. In order to assess the models' accuracy, I conducted an accuracy test by calculating the average of the amount of instances from each model that align with the actual values. The KMeans clustering assigned 67% of the instances correctly, while the agglomerative clustering method assigned 55% correctly.

However, it came to my realization that the models might be overfitting the features. Since all of the DJI stocks are also listed in the SP500, I suspected that the agglomerative model might be overfitting the clusters due to the high levels of similarity in the DJI and SP500 MACDs. So I proceeded to remove the SP500 MACD from the feature set and re-fitted both clustering algorithms. The results of this can be seen in Figures 5. and 6. on the next page.

Figure 5.

Figure 6.

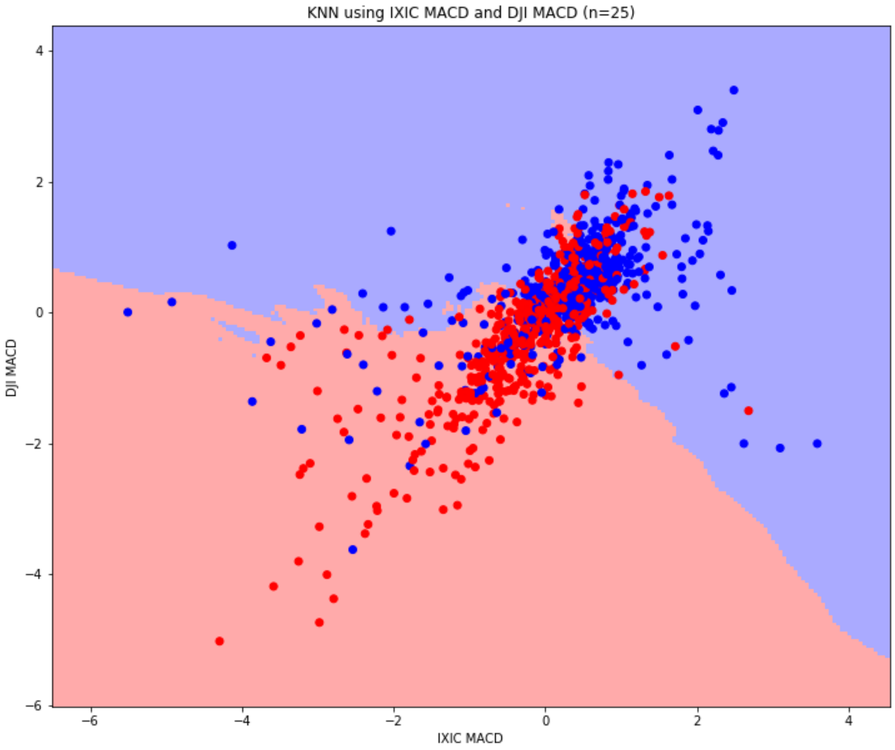


The change in the agglomerative clustering results after removing the SP500 MACD is evident by comparing Figures 2. and 5. It seems to make the separation of the class much more resembling of the actual values. The KMeans clustering algorithm did not seem to change much after the removal of the SP500 MACD. In fact, after this most recent analysis, the agglomerative clustering method seems to be the most visually similar to the actual values. Still, the separation of the clusters is more apparent and distinct when using the KMeans algorithm. I conducted the same accuracy test as I did on the previous clustering analysis. It would seem that the KMeans algorithm has an accuracy of 65%, while the agglomerative algorithm has an accuracy of 58%. Therefore, I am led to believe that the KMeans algorithm is most accurate when clustering the bullish and bearish since its accuracy was greater on two separate occasions.

## Data Analysis: Classification

Since the class values are known to me, I would also like to conduct a classification analysis using the feature set to identify bullish and bearish investor sentiment using the discrete Bull-Bear Spread data. For this analysis I will use the K-nearest-neighbors (KNN) approach with the DJI MACD and the IXIC MACD as my two features in order to stay consistent with my previous clustering analyses. I have fit these two features to the KNN model using *n* = 25. After some trial and error attempts, I found this to be the best *n* neighbors for the classification. The results of the classification can be seen in Figure 7.

Figure 7.



As seen in the visualization above, the KNN algorithm has clearly identified two sections of the plot, separating the bullish from the bearish. The blue area of the plot is expected to be bullish sentiment, while the red area is expected to be bearish. It is clear that there are some inaccuracies since there are multiple instances in either section that do not belong to that classification. I conducted an accuracy for the KNN classification test using the algorithm to predict the values of the instances. The KNN classification predicted 74% of all instances correctly. To further evaluate the classifier, I split the data up into a training and a test set. I fit the training data to the algorithm, and tested its accuracy using the test data set. I repeated this process 10 times in order to train and test the algorithm on 10 different training and testing data set pairs. I found that the average accuracy of the algorithm over these 10 trials to be 72%.

## Data Analysis: Clustering & Classification

Upon completing the clustering and classification analyses, I had the thought to combine the two methods. To do this, I used the same features as the previous classification analysis, but instead of using the actual class values, I used the values produced by the clustering algorithms to fit the KNN model. I then used the applied the model to the actual values of the Bull-Bear Spread in an attempt to classify the bullish and the bearish sentiments. Figures 8. and 9. depict these results.

Figure 8.

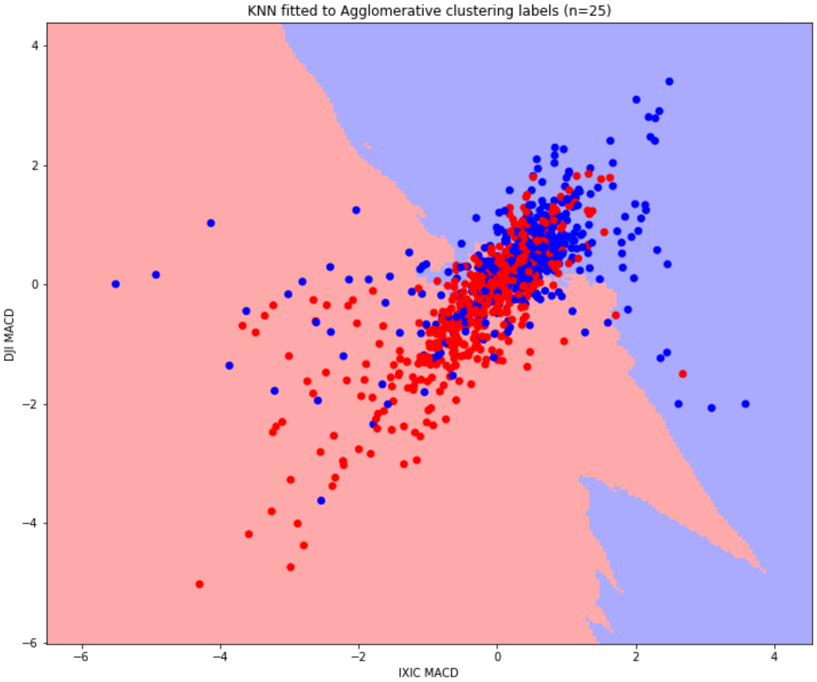
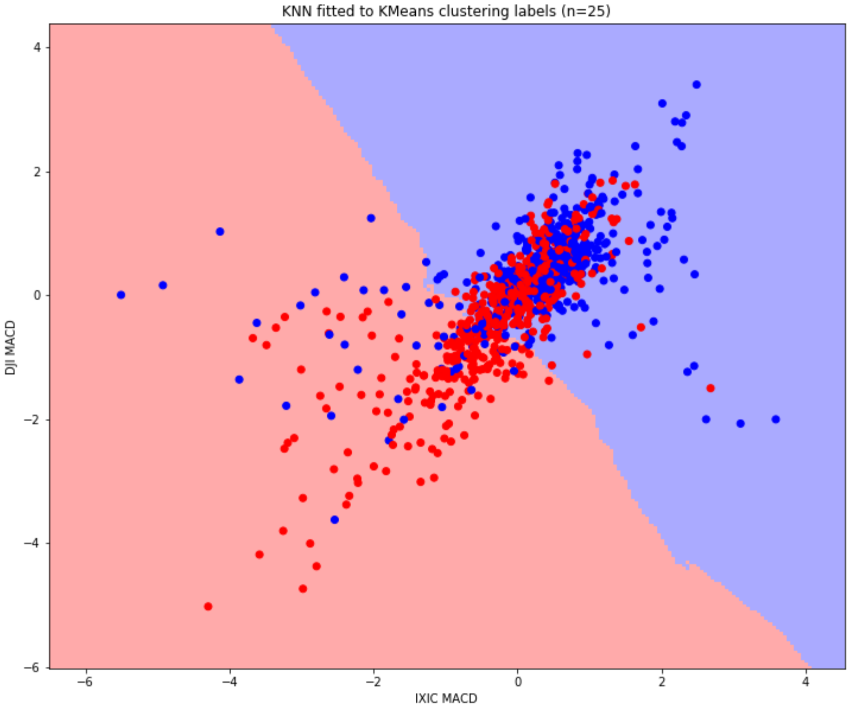


Figure 9.



Although both of these visualizations differ from the original classification visualization seen in Figure 7, they seem to generally classify the bullish values when the MACDs are above 0 and the bearish values when the MACDs are below 0. To assess the accuracy of this method, I conducted the same accuracy test as I did on the original KNN model. When using the agglomerative clustering labels to classify bullish or bearish sentiments, I found that the model accurately classified 66% of the instances. This figure is significantly higher than the accuracy of the clustering model alone. On the other hand, when using the KMeans clustering labels to fit the KNN model, 68% of the instances were classified correctly, which is better than the accuracy of the clustering model alone but not by much.

I conducted another accuracy test using training and testing data similar to that of the original KNN model. I first split the data into training and testing data sets using the same features and the actual Bull-Bear Spread values. However, instead of training the model on the actual class values, I used the clustering labels that were assigned to the class for those instances in the training data. I did so by replacing the class training data with the clustering labels that were in the same index position in the full data set. The purpose of this was to train the data using the clustering labels, but still be able to assess its accuracy using the actual class values as testing data. I repeated this process 10 times, as I had done with the original KNN classification. The results of these trials were consistent with the overall accuracy of the classifications using the clustering labels. When using the agglomerative labels to classify bullish and bearish sentiment, the average accuracy over the 10 trials was 65%, while the average accuracy using the KMeans labels was 69%.

**Conclusions**

The findings of my analysis lead me to believe that the features in my data set do play some role in influencing investor sentiment, and can be used to classify bullish and bearish investor sentiment. The clustering analyses showed some evidence of being able to separate the two sentiments using only the feature set. Although the accuracy of the separation was not high, I would consider the KMeans clustering using all features to be significant since it was able to accurately separate 67% of the data. At random, since this was a binary class, it had a 50% chance of being accurate. Taking this into consideration, I still believe that 67% signifies some level of accuracy other than just random chance. The agglomerative clustering model on the other hand could have been random since the accuracies were just 55% and 57% for the two different feature sets. I do not consider these numbers to be very convincing when compared to random chance.

It is clear that the KNN classification using the class values is the best model to classify bullish and bearish sentiments. An accuracy of 74% is not enough to confidently apply this model to other data sets, but I consider it high enough to suggest that the sentiment behavior changes to some extend depending on the values of the MACDs. This would make sense since they are momentum indicators. If the class values are unknown, the KMeans clustering labels could be used to generate a KNN classifier and then be used to predict the classification. The accuracy for this method was 68%, which gives me some hope for this method.

There are of course many other factors that play into investor sentiment on the stock market than the ones I have included in my analysis. However, a lot of these are difficult to quantify and merge into a time series. For example, the political climate, current trade negotiations and corporate news are very significant factors in investor sentiment, but almost impossible to quantify and then merge to a specific date in the survey data. In order to incorporate some of these factors into my analyses, I could have attempted to conduct a twitter sentiment analysis focusing on politics and corporate news during the weeks in the survey data, and used this sentiment as an additional feature. Unfortunately I did not have the time nor the skills to do so.

**Links to data sources:**

- NASDAQ daily data: https://finance.yahoo.com/quote/%5EIXIC/history?p=%5EIXIC
- S&P 500 daily data: https://finance.yahoo.com/quote/%5EGSPC/
- DJI daily data: https://finance.yahoo.com/quote/%5EDJI?p=%5EDJI
- AAII sentiment data: https://www.quandl.com/data/AAII/AAII_SENTIMENT-AAII-Investor-Sentiment-Data
- US dollar futures data: https://www.quandl.com/data/CHRIS/ICE_DX1-US-Dollar-Index-Futures-Continuous-Contract
- US treasure bond data: https://www.quandl.com/data/CHRIS/CME_US1-U-S-Treasury-Bond-Futures-Continuous-Contract-1-US1-Front-Month

**Calculations sources:**

- EMA: https://www.investopedia.com/terms/e/ema.asp
- MACD: https://www.investopedia.com/terms/m/macd.asp