ISyE 6740 - Summer 2021

Project Report

Firm Strategy and Financial Performance with Natural Language Processing

Erik Magnusson 903661584
Jos Vilier 903638166
Vera Thut 903560685

Contents

1 Problem statement

Firm strategy is big business. There are countless MBA students, financial analysts, and consultants searching for the optimal business strategy. In the following analysis, we employ machine learning techniques to identify firm strategies and quantify their effects on financial performance.

Firm strategy is a result of topics that are top of mind for the firm. There are many ways in which firms try to achieve competitive advantage depending on what they believe are the most important strategic factors. For example:

- Firms may choose to "go green" if they are concerned about climate change
- Firms can focus on differentiation if they believe consumer behavior is changing
- Firms can focus on cutting costs if they deem their products interchangeable with their competitors

In this project, we measure how a firm's performance is affected by its strategic choices. The textbook answer would be yes: a firm that makes good strategic choices will be able to generate more revenues and/or reduce costs, leading to additional profits. These additional profits would be reflected in a higher stock price.

2 Data sources

The two main data inputs for this analysis are firm strategy and financial performance, collected as text files and time series, respectively.

The analysis is restricted to firms in the consumer packaged goods (CPG) food industry. Because of the common consumer-facing nature of this sector, many people are familiar with these companies and can understand their business structures and strategies with relative ease. It is also a large sector with many publicly traded companies, providing a large selection of available annual reports for analysis. The CPG companies used in this analysis are restricted to firms in the S&P 500 Index to ensure that the components of the holding period return are freely available.

2.1 Firm strategy

Firm strategy is particularly hard to capture, as it is not a quantitative variable that firms are required to publish. However, all publicly traded companies in the United States are required to file their annual reports with the US Securities and Exchange Commission (SEC). The SEC stores this information in their database EDGAR.

The reports follow a standard template. Most information is accounting related, e.g., a balance sheet and profit and loss statement over the calendar year. Each report is largely backward-looking, focusing on the company's past performance. However, every report contains a Risk Factors section that is always forward-looking. This is where a firm identifies and manages risks that may threaten their business in the short and long term. It generally includes subjects such as increased regulation, climate change, changing consumer preferences and potential litigation. It contains valuable information about the topics that are top of mind at the firm and may shape firm strategy. By extracting the text in these risk sections, we generated a data set for identifying patterns regarding firm strategy.

The qualitative parts of annual reports (such as the risk factor section) are generally written in November and December of the calendar year. For example, the 2018 annual reports (containing profits and losses for the year 2018) tends to be published in March or April 2019, but the risk factor section was written in November 2018 and focuses on potential risks in 2019 and beyond. For this reason we are going to link the annual report over 2018 to the financial performance over 2019. We also choose to exclude the year 2020 from the data set, because the pandemic would lead to many outliers. Furthermore, to reduce the influence of autocorrelation we do not include consecutive years. We chose to use one in every three years for this reason. Going back further in time may lead to data availability issues, hence we restrict ourselves to 2012 and further. Table 1 contains what years are selected.

We used annual reports from 14 companies and have total of 39 observations in our data set. The annual reports were uploaded from AnnualReports in PDF format. As formatting in the annual reports varies, the Risk Factors data was manually extracted from each of the report and converted to individual .txt files.

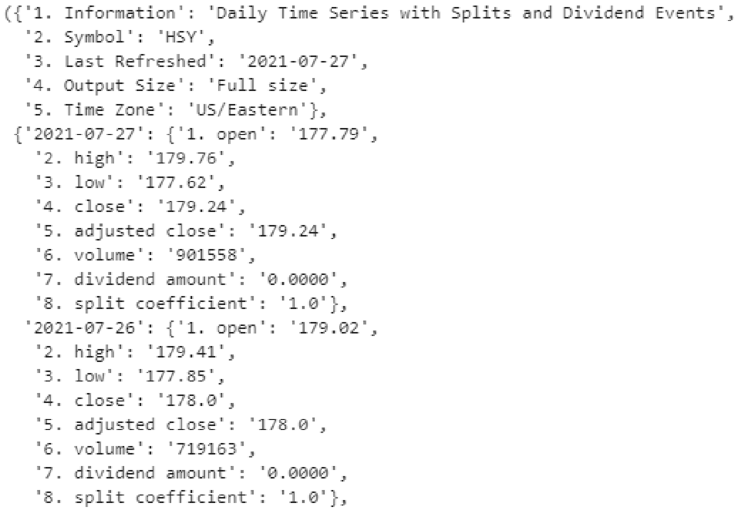
Cohort	**Annual report**	**Return**
1	2012	2013
2	2015	2016
3	2018	2019

Table 1: Years in the data set

2.2 Financial performance

Unlike firm strategy, financial performance is relatively easy to quantify. Financial performance is calculated using the difference between the annual return of a given public company, adjusted for dividends and events like stock splits, and the overall annual return of the S&P 500 Index. This calculation generates the company's excess returns for a given year and is used as the benchmark for financial performance.

Each company's financial performance is captured using the Alpha Vantage API, which returns JSON-formatted time series stock data. This contains historical daily close prices for public companies, as well as any relevant stock split/dividend event data. By executing an API call for each of the consumer packaged goods companies of interest, we obtain 20+ years of historical financial performance data.



```
({'1. Information': 'Daily Time Series with Splits and Dividend Events',
  '2. Symbol': 'HSY',
  '3. Last Refreshed': '2021-07-27',
  '4. Output Size': 'Full size',
  '5. Time Zone': 'US/Eastern'},
 {'2021-07-27': {'1. open': '177.79',
   '2. high': '179.76',
   '3. low': '177.62',
   '4. close': '179.24',
   '5. adjusted close': '179.24',
   '6. volume': '901558',
   '7. dividend amount': '0.0000',
   '8. split coefficient': '1.0'},
  '2021-07-26': {'1. open': '179.02',
   '2. high': '179.41',
   '3. low': '177.85',
   '4. close': '178.0',
   '5. adjusted close': '178.0',
   '6. volume': '719163',
   '7. dividend amount': '0.0000',
   '8. split coefficient': '1.0'},
```

Figure 1: Alpha Vantage API output: Hershey (HSY) example

Historical S&P 500 Index returns are available through Berkshire Hathaway and are assumed to represent an estimate of overall market return for a given year. Subtracting these market returns from the corresponding adjusted annual returns for each CPG company yields the financial performance benchmark for this analysis. Because annual reports from the years 2012, 2015, and 2018 are being considered, excess returns for the years 2013, 2016, and 2019 are calculated.

3 Methodology

The methodology consists of three steps. There are two options for the third step, which will both be explored to ensure robustness of the results. The steps are explained in more detail below.

3.1 Natural language processing

In this step we convert the risk factor section in the annual reports to a vector that captures the topics that are top of mind at the firm. We do this by applying a bag of words model, which simply consists of counting how often each word is used. There will likely be a step in this process where we specify what words should be considered (i.e. what are the features). For example, the word ”is” will likely be used very often, but it does not convey information regarding firm strategy. These words will be removed from the feature vector. This step is performed manually and can be rather subjective. The selected and removed words are available in ”/data/bag of words/bow_unfiltered_tagged.xlsx”. 422 out of 5,651 unique words are selected. Note that so called stop words are filtered for as well.

It is important to realize that the total number of words can be very different from one annual report to the next, hence we correct for this by standardizing with the overall word count per annual report.

The result of this step is a matrix that contains the relative word count for each observation (row) for every considered word (column). The rows sum up to 1. An example is given in table 2. The size of the matrix is 42 (firms) by 422 (unique words related to strategy). The matrix is rather sparse.

Firm & Year	**Accountants**	**Accounting**	**Acquire**	**...**	**Workforce**
ADM_2012	0	0.005	0	...	0.005
ADM_2015	0	0.004	0	...	0.004
ADM_2018	0	0.003	0	...	0.003
...	...	...	...	...	...
TR_2018	0	0.006	0	...	0

Table 2: Bag of words results

Over all of the annual reports in the data set, the 20 most common words, measured by the sum of their relative frequencies, are shown in Figure 2. Many of the most frequently used words could be considered generic commercial terminology, such as ”business”, ”financial”, and ”customers”. However, the manner in which an organization stresses these different components could provide insight into broader strategic intent.

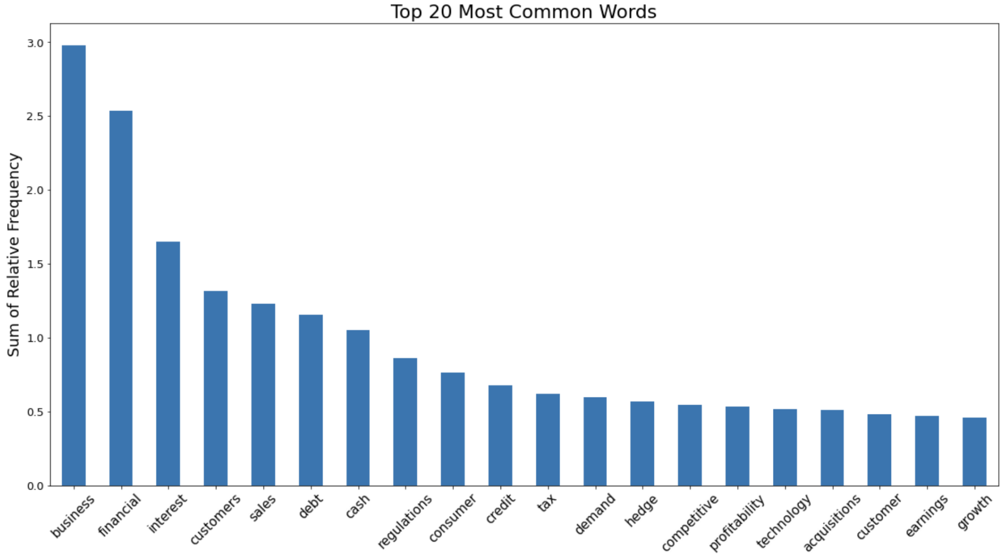
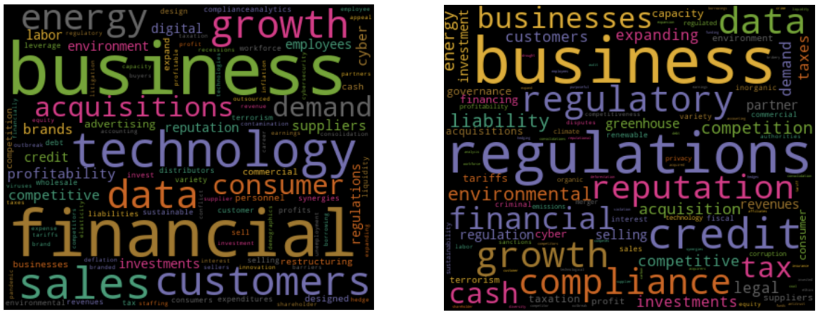


Figure 2: Most common words, by relative frequency, in annual report risk factor sections

For example, Figure 3 shows two word clouds from example annual reports in the data set. Common terminology like "business", "financial", "growth", and "data" are mentioned frequently in both the HSY and ADM 2018 annual reports. However, in addition to the high usage of those shared terms, each company differentiates its strategic focus by repeatedly using more specific terminology. For example, HSY has higher relative word frequencies for "customers", "sales", "energy", and "technology", while ADM has higher relative word frequencies for "regulations", "compliance", "credit", and "reputation".



(a) HSY 2018 Annual Report

(b) ADM 2018 Annual Report

Figure 3: Example word frequencies for Risk Factors sections

3.2 Dimensionality reduction

The number of word frequency features generated by the bag of words analysis (n = 422) is much greater than the number of observations (m = 42) in the data set. To avoid having more predictors than observations, which can lead to issues like overfitting, principal component analysis (PCA) is used to reduce dimensionality. This is accomplished by projecting the original data set onto the first few principal components, belonging to an orthonormal basis of the original feature space, in an effort to reduce the number of features while maximizing the amount of data variation retained.

PCA allows us to recognize which word combinations account for the most variation between observations and provides insights into the strategies being discussed in each annual report. Figure 4 displays the proportion of overall variance from the original data set that is explained by each subsequent principal component. Given the dimensions of the data set, the most effective modeling will come from minimizing the number of principle components used while maximizing their cumulative explained variance.

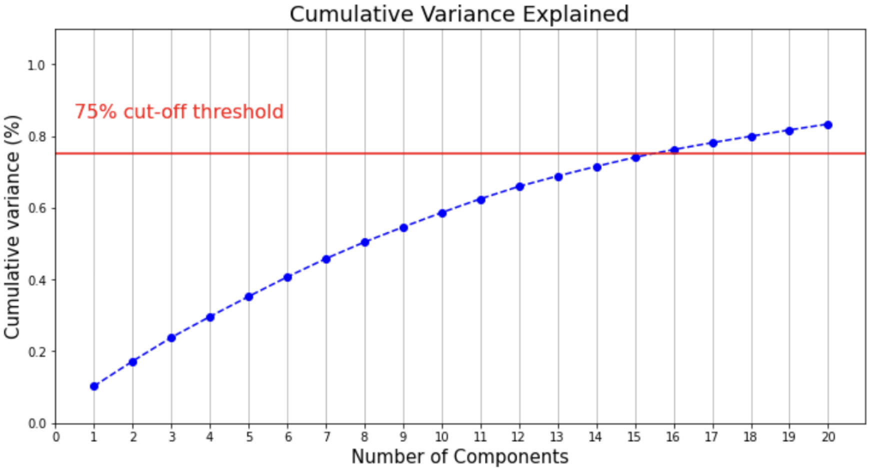


Figure 4: Cumulative variance explained by the first 20 principle components

While it would be ideal to select a number of principle components that explains a cumulative variance percentage closer to the arbitrary 75% threshold shown in Figure 4, it would be counterproductive given the small annual report sample size. Also, as seen in Figure 5, there is no sudden drop off in explained variance between the different components, creating a situation in which there is no natural number of components that should be considered. This leads to a bias-variance trade-off which depends on the number of principle components included in the analysis. Including more principle components allows more of the original variance in the data to be incorporated in the modeling process. However, introducing too many features relative to the number of observations can lead to increased variance when generalizing to new data.

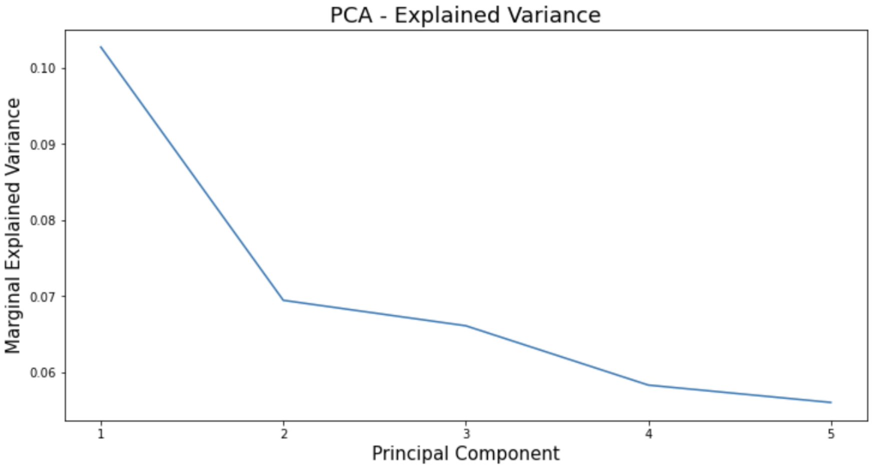


Figure 5: Marginal variance explained by the first 5 principal components

We decided to go with five components to be able to capture a wide range of strategies. Table 3 contains the five words associated with the largest (positive) and smallest (negative) values for each component.

	Component 1	**Component 2**	**Component 3**	**Component 4**	**Component 5**
Positive	Noncompliance Pollution Afford Commercialization Compliant	Acquired Brand Brands Patent Consumers	Reputation Regulatory Selling Cyber Invested	Analysts Affordably Analyst Appealing Budgets	Expenditures Acquisitions Profitability Pandemic Demographic
Negative	Energy Environment Competitive Pandemic Cyber	Energy Demand Outbreak Hedge Terrorism	Debt Cash Acquire Interest Earnings	Sellers Distributors Buyers Employee Innovation	Interest Hedged Hedge Accounting Bribery

Table 3: Principal component analysis results

These words allow for interpretation of the principal components. For example, the words ”energy”, ”environment”, ”competitive”, ”pandemic” and ”cyber” are commonly used in reports that have a low value for the first component. This indicates that the firm is concerned about catastrophes because all words except ”competitive” are associated with catastrophes. Positive values for the first component indicate that they are concerned about compliance issues e.g. with food and drug authority rulings (”noncompliance”, ”compliant”) and commercial aspects (”afford”, ”commercialization”). The components have been analyzed and translated into strategies in table 4. The strategy indicates what the firm focuses on.

This step is rather subjective. There are a number of ways in which this analysis can be improved, e.g. by correcting for synonyms, by collecting more data or by considering more than the top five words to decide on the strategy. All these approaches require substantial time investments however.

Component	Positive	Negative
1	Compliance issues and commercial aspects	Catastrophes
2	Marketing and consumer needs	Catastrophes
3	Compliance issues	Financial aspects
4	Consumer needs	Competitive environment
5	Mergers and acquisitions	Financial aspects

Table 4: Component interpretation

3.3 Classification and clustering

In this section we are going to test in two ways whether there is any link between firm strategy and financial performance. The following two approaches both start after the PCA step, hence the project branches off in two models. We apply clustering and classification to analyze the results.

3.3.1 Clustering

Approach number one starts with clustering of the observations. Next we are going to compute the average outperformance of the S&P 500 for each cluster and compare these. Large differences between this outperformance are indicators that strategy indeed plays a role to explain financial performance. The clustering is done by the well known k-means model based on Euclidean distances. The PCA data is standardized before applying the model.

The sum of squared distances (also known as residuals) are plotted as a function of the number of cluster centers in figure 6.

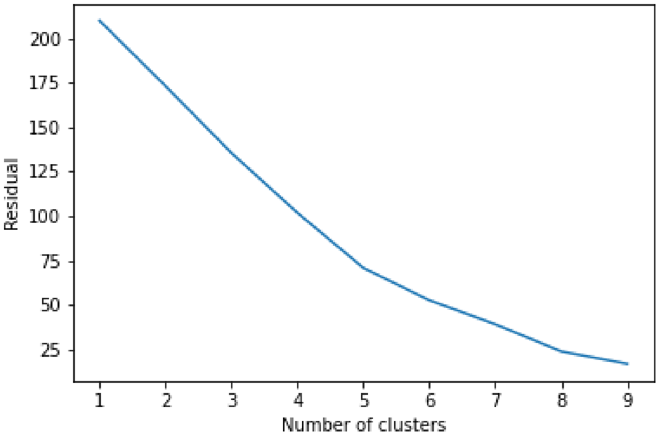


Figure 6: Residuals for the k-means model

The graph does not show a clear elbow point. The most likely candidates are k = 5 and k = 8, however those are rather high cluster numbers with only 42 observations. We decided to go with k = 3, which leads to the clustering in table 5 and cluster centers in table 6. Note that we did not have three years of history for ANFI and NOMD.

The clustering shows that ANFI is so different from the other reports that it gets it own cluster. In addition, the clustering seems to capture that there is autocorrelation between the different years, as almost all firms remain in their cluster (GIS is the only exception). Cluster 1 and 2 are both reasonably populated compared to cluster 0.

The cluster centers for cluster 1 and 2 are not that different. It is mainly component 2 where the centers are quite far apart. Resorting back to the strategies in table 4 shows that cluster 1 mainly focuses on catastrophes (as the value of component 2 is negative), while cluster two focuses on marketing and consumer

Cluster 0	**Cluster 1**	**Cluster 2**
ANFI_2015	ADM_2012	BGS_2012
ANFI_2018	ADM_2015	BGS_2015
	ADM_2018	BGS_2018
	HSY_2012	CAG_2012
	HSY_2015	CAG_2015
	HSY_2018	CAG_2018
	INGR_2012	CPB_2012
	INGR_2015	CPB_2015
	INGR_2018	CPB_2018
	SJM_2012	K_2012
	SJM_2015	K_2015
	SJM_2018	K_2018
	TR_2012	MKC_2012
	TR_2015	MKC_2015
	TR_2018	MKC_2018
	GIS_2012	GIS_2018
	GIS_2015	NOMD_2018
		POST_2012
		POST_2015
		POST_2018
		THS_2012
		THS_2015
		THS_2018

Table 5: Clusters for k = 3

Cluster	**Component 1**	**Component 2**	**Component 3**	**Component 4**	**Component 5**
0	3.97	-1.19	0.39	-1.07	0.02
1	-0.47	-0.97	0.18	0.25	-0.05
2	0.00	0.82	-0.16	-0.09	0.04

Table 6: Cluster centers for k = 3

needs (as the value of component 2 is positive). Cluster 0 stands out in the first component. The results for cluster 0 are not that interesting however, as it seems to capture an outlier only.

The average outperformance for each cluster are -40.7%, -0.6% and -3.7%. This shows that the CGP food industry performed worse than the S&P500 in general, but cluster 1 had a stronger performance compared to cluster 2. This seems to indicate that it is better to focus on catastrophes than marketing and consumer needs. However, the performance may also be (subtly) influenced by the other components (or strategies), hence we will take a closer look at all components in the following section on classification models.

3.3.2 Classification

For our analysis, we used a number of supervised learning algorithms to understand whether features can predict outperformance (return higher than S&P 500) or underperformance (return lower than S&P 500) of a particular company.

Supervised Learning models used:

- **Naive Bayes**.
  Strengths: This model can deal with many features. It is simple and fast.
  Limitations: classifier assumes that all attributes are mutually independent.
- **KNN**.
  Strengths: This algorithm is easy to implement and rather versatile.
  Limitations: it can get slow, as the number of features increases.
- **Logistic Regression**.
  Strengths: This classifier is easy to implement, it is also less prone to over-fitting.
  Limitations: it is hard to understand complex relationships using this algorithm.
- **Linear SVM**
  Strengths: Effective for high-dimensional data sets. Memory efficient.
  Limitations: Does not perform well for data set with a lot of noise. Can under perform if number of features exceeds number of training data samples.

Out or under performing S&P 500 was defined as a Target Variable. The data was split 70/30 for train/test. The classifiers were run using the calculated principle components as features. We experimented utilizing varying numbers of principle components: 2, 3 and 5. Below are the accuracy scores for each classifier with each different set of principle components:

Data	Naive Bayes	KNN	Logistic Regression	Linear SVM
5 principle components	0.55	0.4	0.4	0.4
3 principle components	0.6	0.5	0.5	0.45
2 principle components	0.4	0.6	0.55	0.5

Table 7: Accuracy classification scores

Further, we created plots to understand if any of the combinations of 2 components within the **3 principle component** set would perform better. The results are shown below:

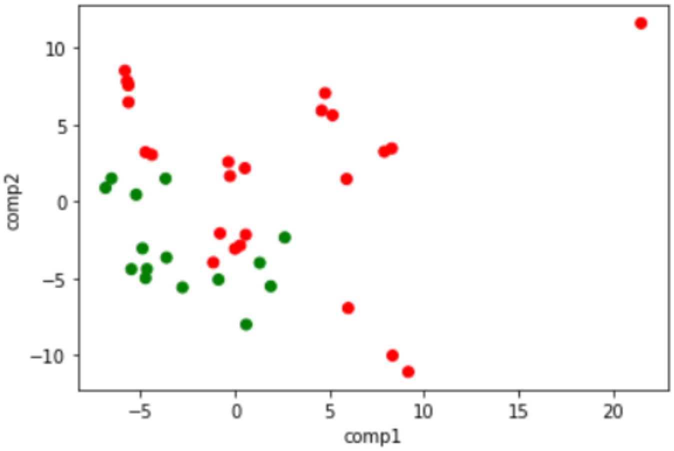


Figure 7: Plotting Principle Components 1 and 2

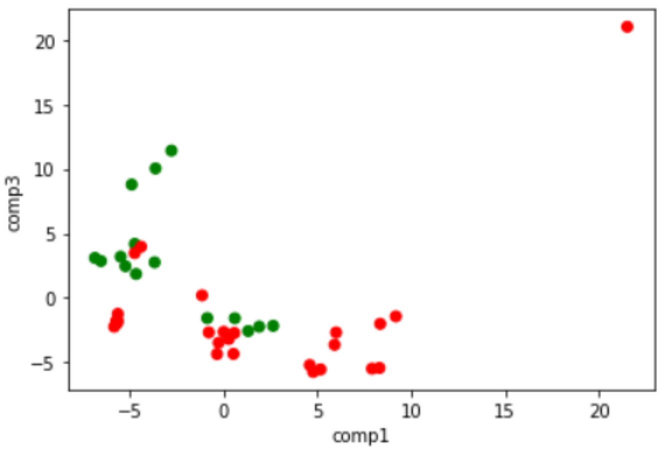


Figure 8: Plotting Principle Components 3 and 1

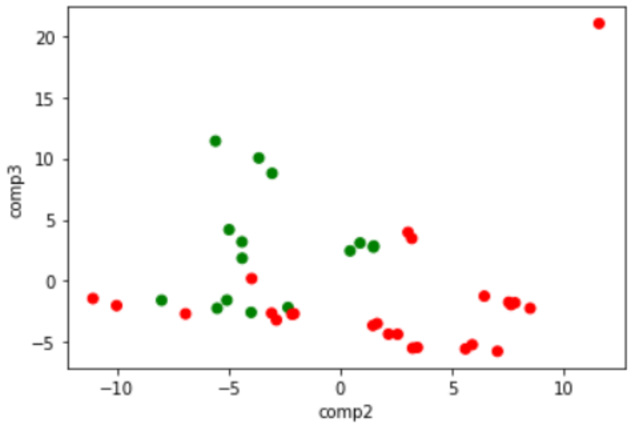


Figure 9: Plotting Principle Components 3 and 2

As we see from the charts above, the combination of Component 1 and Component 2 yields the best result. We also can see that outperforming data points are more associated with lower values for both Component 1 and Component 2, and thus can be tied to the negative words in those components, like "Energy", "Environment", "Competitive", "Demand", "Outbreak", etc..

All classifiers are then run using 5, 2 out of 3, and 2 principle components, looking at all metrics and assessing the performance of each of the classifiers.

Data	Accuracy score	Recall Scores	Precision Scores	F1 Scores
PC 1 and 2 out of 3	0.58	1., 0.25	0.73, 1.	0.84, 0.4
2 PC	0.58	0.63, 0.5	0.71, 0.4	0.67, 0.44
5 PC	0.42	0.6, 0.29	0.38, 0.5	0.46, 0.36

Table 8: Naive Bayes

Data	Accuracy score	Recall Scores	Precision Scores	F1 Scores
PC 1 and 2 out of 3	0.5	0.5, 0.5	0.67, 0.33	0.57, 0.4
2 PC	0.67	0.5, 1.	1, 0.5	0.67, 0.67
5 PC	0.42	0.8, 0.14	0.4, 0.5	0.53, 0.22

Table 9: KNN

Data	Accuracy score	Recall Scores	Precision Scores	F1 Scores
PC 1 and 2 out of 3	0.5	0.63, 0.25	0.63, 0.25	0.63, 0.25
2 PC	0.17	1, 0	0.83, 0	0.91, 0.
5 PC	0.5	0.6, 0.43	0.43, 0.6	0.5, 0.5

Table 10: Logistic Regression

Data	Accuracy score	Recall Scores	Precision Scores	F1 Scores
PC 1 and 2 out of 3	0.5	0.5, 0.5	0.67, 0.33	0.57, 0.4
2 PC	0.58	1., 0.	0.58, 0.	0.74, 0.
5 PC	0.33	0.6, 0.14	0.33, 0.33	0.43, 0.2

Table 11: Linear SVM

The results in tables 8, 9, 10, and 11 show that most of data/classifier combinations produce relatively low accuracy scores around or below 0.5. The highest accuracy score of 0.67 was achieved with a KNN classifier based on two principle components. The F1 Scores are relatively high for this classifier, however Recall and Precision Scores are not entirely conclusive. Overall, the better performance of KNN can be explained by the fact that this classifier works the best for non-linear data relationships (like stock market data). The selection of two principle components for this model is not arbitrary because, as it is stated in the dimensionality reduction section of the report, there is no clear cut off for selecting the number of principle component and theoretically, including more principle components would allow us more of the original variance in the data to be incorporated.

Overall, for classification models the results of the experiments are inconclusive and need to be investigated further. Yet, several adjustments are recommended to be made:

- Our data set is relatively small; we have around 20 instances of each class. Although such algorithms as Naive Bayes and KNN can work well with smaller data sets, the current results are inconclusive and need to be verified on a larger data set. We can expand it by adding results for more years, including annual reports of more companies or both.

- Neural Network classification can be used for classifying initial ”bag of words” data sets into ”classes”, helping to mitigate bias introduced by filtering/reducing the subset of words manually.

4 Evaluation and final results

The interpretation of the classification model results is critical to answer our research question. Unfortunately, the performance (in terms of accuracy) of all the models is not as strong as we would like. The remainder of this section focuses on the models with five principal components, as these also provide information on more strategies (e.g. mergers and acquisitions) compared to the two or three component models.

Our feature vector consists of five dimensions, which makes it impossible to plot directly in a 2 dimensional report. Figures 10 and further help with this interpretation. Each plot contains of a grid of decision boundaries for two principal components at a time. The remaining three variables are kept constant at their mean value (which is zero because of the PCA procedure).

Each grid is five by five, reflecting the five principal components that we considered in this step. The diagonal is empty; these would show decision boundaries where the x and y axis represent the same component, which is not in line with how decision boundaries are generally interpreted. The off-diagonal entries show the decision boundaries where the component on the x axis is given by the row number and the component on the y axis is given by the column number (which is labeled above the column). The following examples are based on the Naive Bayes classifications in Figure 10. For example, the bottom-left graph contains component 5 on the x-axis and component 1 on the y-axis. The top-right graph is exactly the same except for a switch in axes (i.e. component 1 is represented by the x axis and component 5 the y axis).

Going back to the bottom left graph, we can see that low values for component 1 (y axis) lead to underperformance (red). High values for this component indicate outperformance. We can judge the overall impact of a component by looking at the y axis under its label. For example, for component 1 we see that the top of the graph always tends to be ”more green” than the bottom, hence high values for component 1 are indicative of outperformance. Some columns do not give a clear picture, such as those for component 2 and 4.

It is important to remember that these visualizations are two dimensional slices through the origin of a five dimensional space. There may be other parts in the five dimensional space where the classification behaves differently. For example, the graph for the Naive Bayesian model clearly depicts the ellipsoid shape of the decision boundary in the first graph under component 4.

In the next step, we decide for each component and model whether higher or lower values for the component lead to outperformance. Some of these remain undecided because of conflicting (or no) decision boundaries. For example, the KNN classification leads to very noisy predictions. Table 12 contains the overall outcome based on all the graphs.

Model	**Component 1**	**Component 2**	**Component 3**	**Component 4**	**Component 5**
Naive Bayes	Higher	Undecided	Higher	Undecided	Undecided
KNN	Undecided	Lower	Undecided	Undecided	Undecided
Logistic Regression	Higher	Lower	Higher	Lower	Higher
Linear SVM	Higher	Higher	Higher	Lower	Higher
Overall	Higher	Undecided	Higher	Lower	Higher

Table 12: Classification model results

The most likely strategy for outperformance has high values for components 1, 3 and 5, and low values for component 4. Reverting back to table 4 shows that this corresponds with a strategy that focuses on compliance issues, commercial aspects, the competitive environment, and mergers and acquisitions. The strategy should avoid focusing on catastrophes, financial aspects, and consumer needs.

As the accuracy of the classification models were not particularly high, we can only treat this as a preliminary conclusion. There are several ways in which the analysis can be improved, namely:

- The number of companies and years included in the data set can be increased to get more reliable results

- Models that take autocorrelation explicitly into account can be considered for more accurate modeling of text and returns over time

- The keyword selection in the bag of words model can be improved upon by starting with a defined set of words
- The bag of words model can be improved by also taking synonyms, similarity, and negation into account
- The dimensionality reduction step can be improved by trying other (non-linear) to capture more of the variability with less components
- The clustering section can be improved by considering different models or different distance metrics
- Classification models may be improved with further hyperparameter tuning (e.g. the number of neighbors in KNN)
- The translation from keywords to strategy should be done in cooperation with a (panel of) business strategy expert(s) to ensure that these are all viable strategies

Successful implementation of these steps may increase the accuracy of the classification models and our conviction of the (potentially new) conclusions.

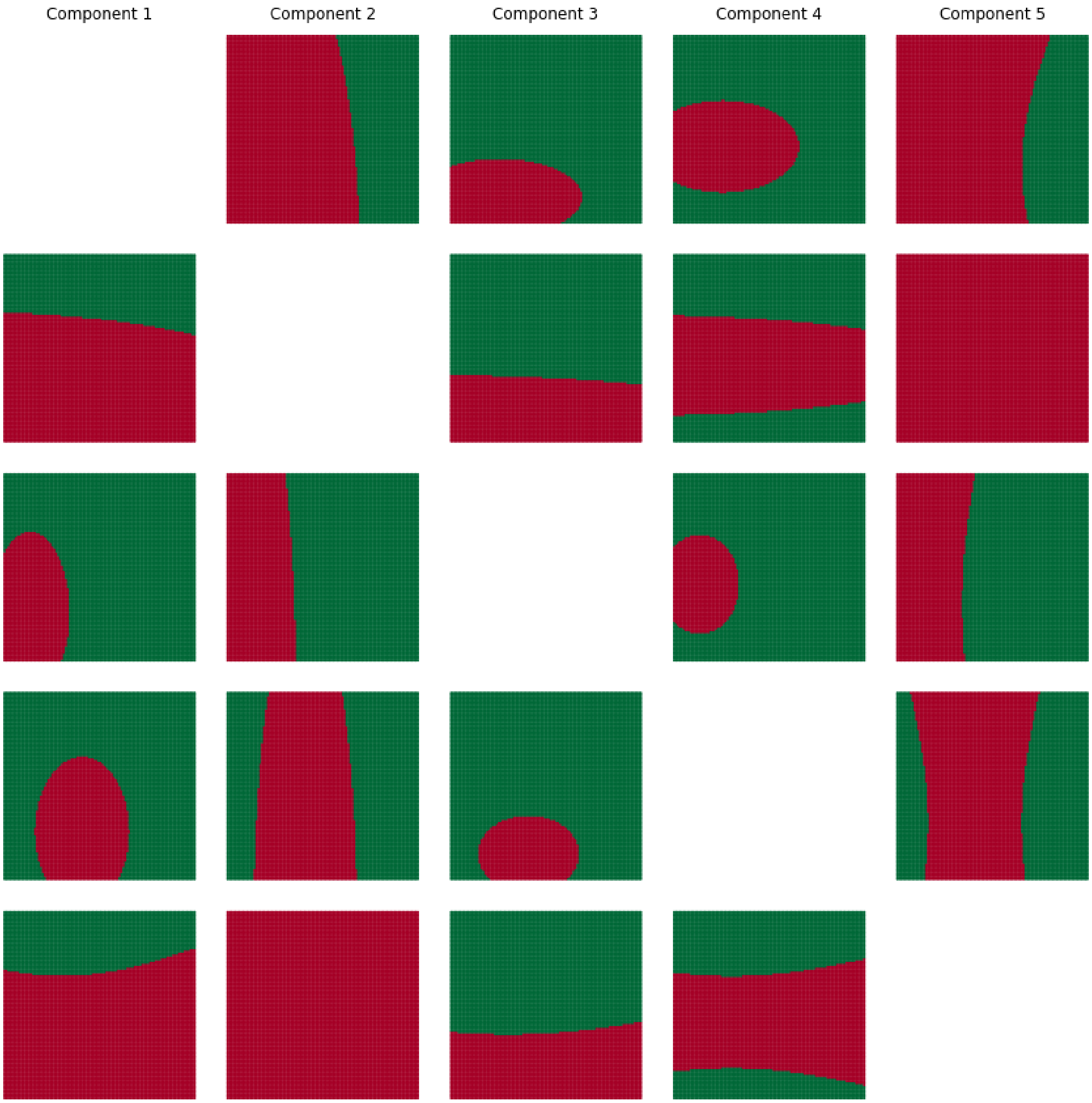


Figure 10: Decision Boundaries Naive Bayes

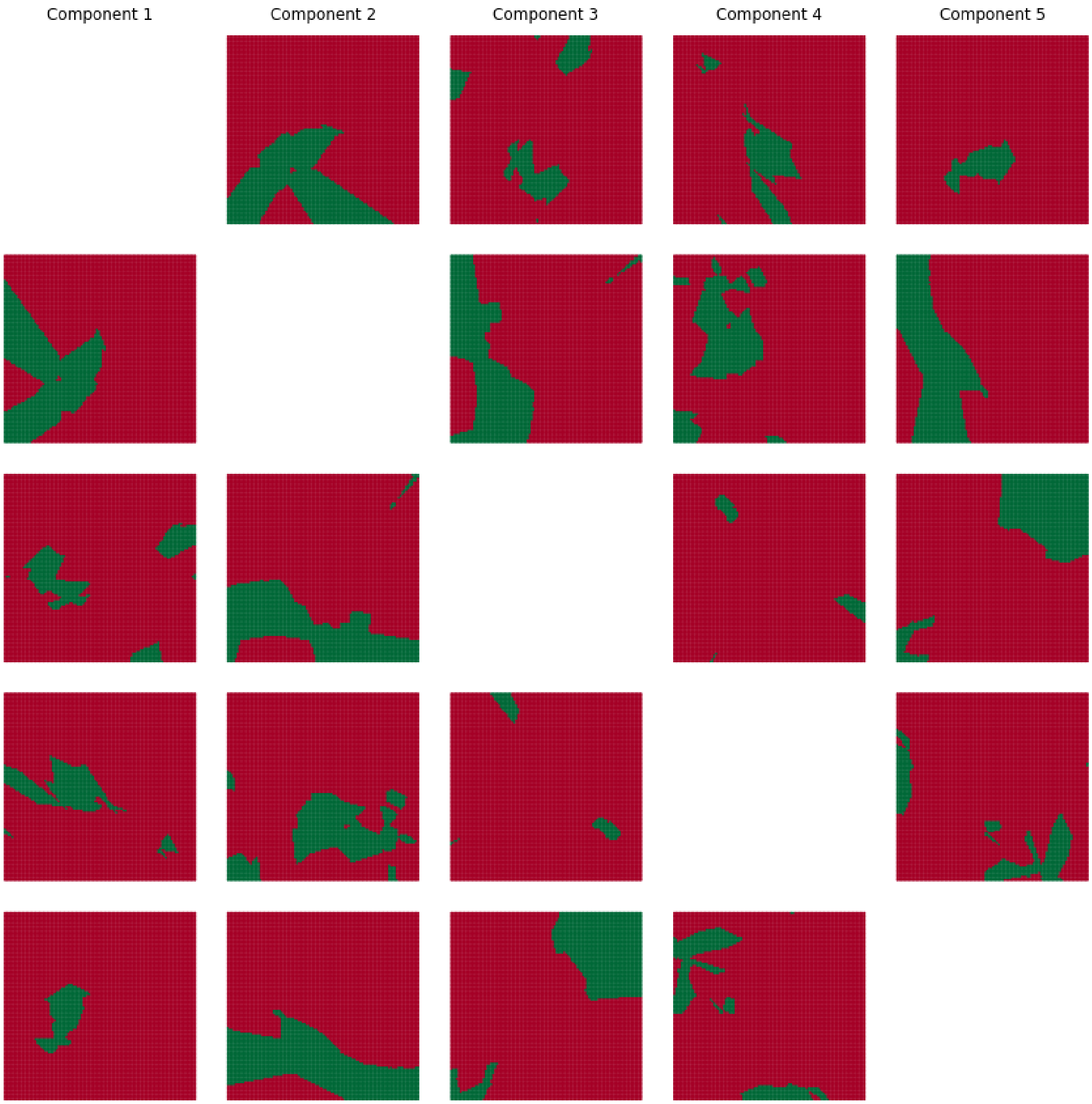


Figure 11: Decision Boundaries KNN

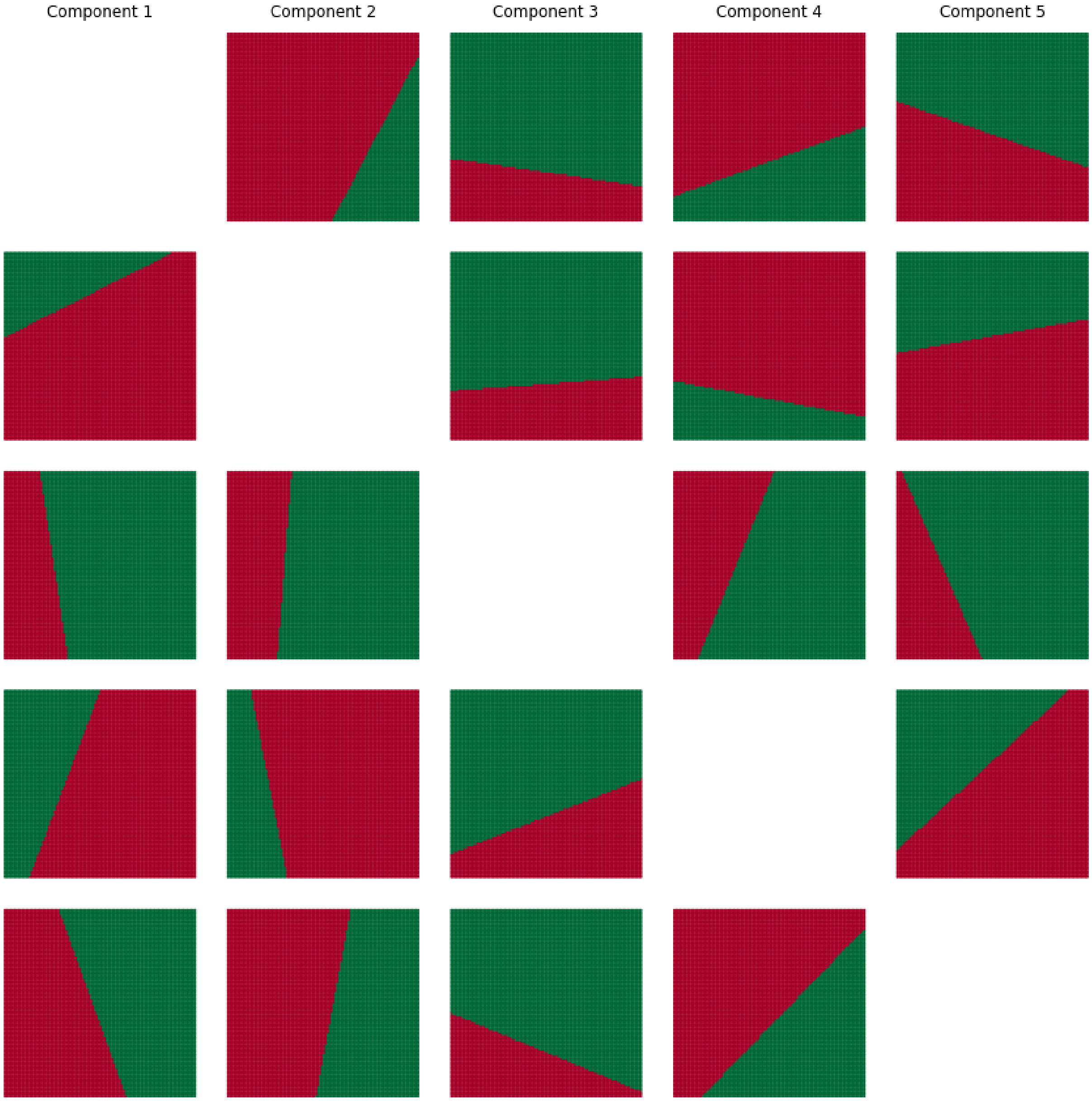


Figure 12: Decision Boundaries Logistic Regression

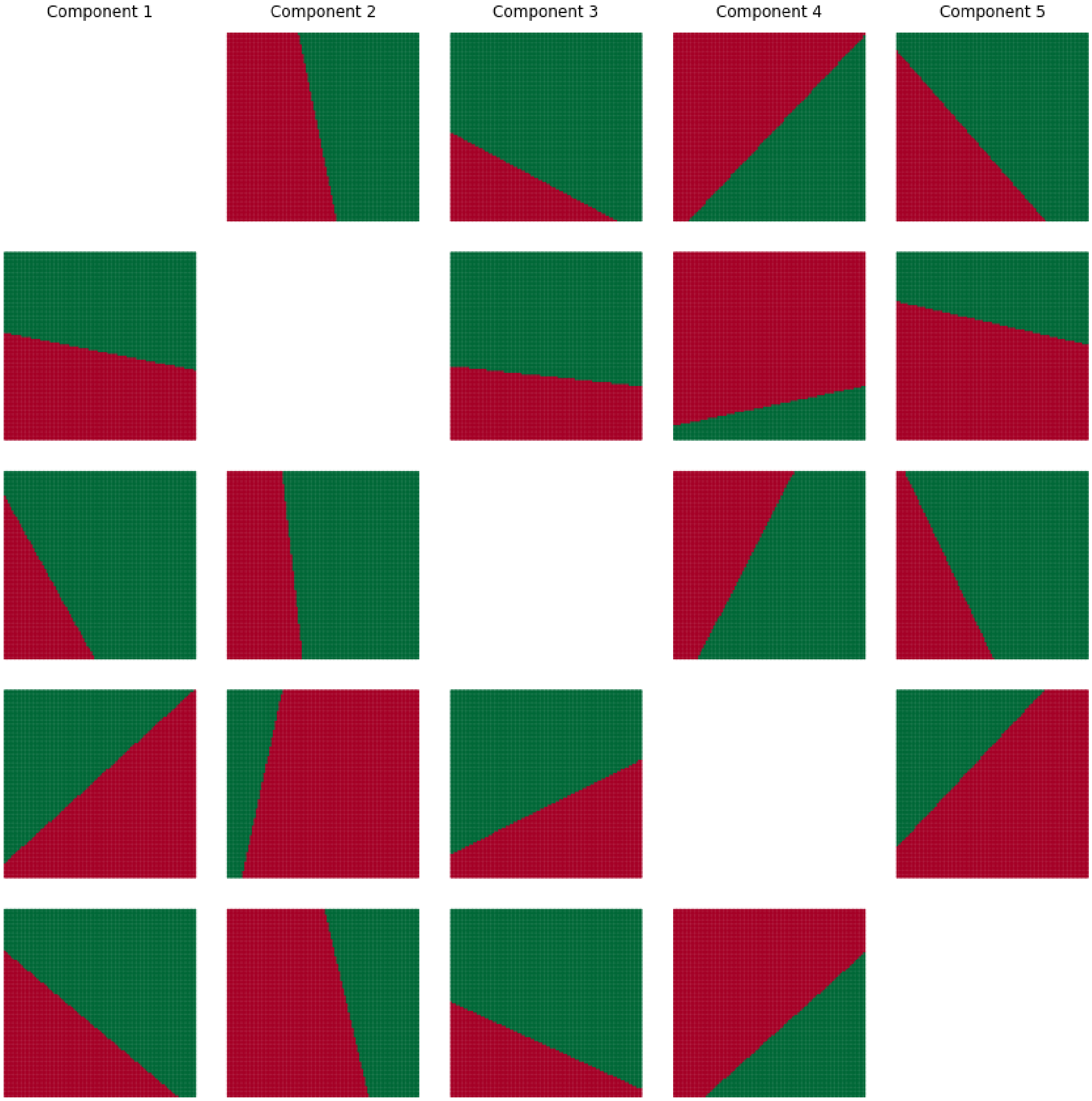


Figure 13: Linear SVM Decision Boundaries

Appendix A: Who did what

This section contains an overview of all the tasks the individual team members performed, in line with project requirements.

Erik

- General
  - Using Alpha Vantage API and Berkshire Hathaway reporting to gather/transform financial data
  - Code clean-up/reformatting for python notebook submissions
- Report
  - Section 2.2 Financial performance
  - Section 3.1 Natural language processing (second half)
  - Section 3.2 Dimensionality reduction (first half)
- Coding
  - exploratory_data_analysis.ipynb: exploring the annual report text data
  - pca.ipynb: conducting principal component analysis
  - stock_data_import.ipynb: collecting financial data from API

Jos

- General
  - Initial idea, leading and taking minutes of project meetings
  - Keyword selection (bringing 5,651 unique words back to 422 business keywords)
  - Wrote proposal
- Report
  - Section 1 Problem statement
  - Section 2.1 Firm strategy
  - Section 3.1 Natural language processing (first half)
  - Section 3.2 Dimensionality reduction (second half)
  - Section 3.3.1 Clustering
  - Section 4 Evaluation and final results
- Coding
  - read_txt.ipynb: reads text files and returns bag of words model
  - clustering.ipynb: performs cluster analysis on five components
  - Code to create the 5 by 5 decision bounds visualization

Vera

- General
  - Searching and downloading annual reports from Annual Reports
  - Extracting data from annual reports and saving data in separate .txt files

- Report
  - Paragraph on data sourcing in Firm Strategy section
  - Full report on Classification in Classification and Clustering section
- Coding
  - File PCA + Models_FINAL.ipynb (some PCA related code + Decision Boundaries code was prepared by Erik and Jos correspondingly, specific cells are commented in the Notebook)