Unsupervised Learning to Confirm S&P 500 Sectors

Atish Sawant
New York, USA
atishsawant87@gmail.com

***Abstract*—This study seeks to identify sectors within the S&P 500 given each firm's 10-K report using NLP and clustering algorithms. The quarterly reports filed by each firm identify the firm's performance, risks, and operating concerns. Thus, if clusters are found and they correlate strongly to the existing stock market sectors then it implies that superior risk management and performance can be found by using news headlines and articles to identify which stocks that that article most closely relates to.**

I. Introduction

The S&P 500 is the preferred index for equity investors around the globe when analyzing the strength and returns of the U.S. market. Currently there are 11 sub-sectors that this market is broken down into, with the 11th sector, Telecommunications, containing just AT&T and Verizon. Every 3 months, every publicly listed company in the United States must submit a report to the SEC detailing their performance, risk measures, and operating characteristics. This report known as the 10-Q, and the annual version known as the 10-K are widely seen as the most important documents in determining the quality and direction of any firm.

Using only these reports, this study uses NLP techniques to reduce the reports into a bag of words in order to perform unsupervised learning using several clustering techniques. The algorithms this study employs are LDA, NMF, and K-Means. For LDA the documents are reduced to word count matrices with stop words, numbers, and all special characters removed. For K-Means and NMF, TF-IDF was used to convert the documents into matrices, and once again stop words, numbers, and special characters were removed.

The goal of the study is to identify clusters of topics using just the reports submitted to the SEC, and quantify how closely these sectors line-up with the established ones. The intuition behind this approach is that firms that mention the same sort of risks, operating metrics, and business results likely are correlated in some way. Ultimately, if strong clusters are found, headlines and news article contents can be associated with clusters. If investors were able to quickly identify which cluster of stocks a particular article impacted, it would allow for superior risk management and enhanced returns.

II. Related Work

Natural Language Processing (NLP) has been a very active field of research in recent years. Breakthroughs such as word2vec, LSTM application to text generation, and semantic analysis have all had a large impact on the field. The use of NLP within financial markets has largely been to directly predict stock market returns using sentiment analysis or documents.

Prior attempts at clustering financial data has mostly been related to matching the similarity of returns. Kevin A.J. Doherty, Rod G. Adams, Neil Davey, and Wanida Pensuwon previously succeeded in clustering stocks in the FTSE based on the pattern of daily returns from 1995-2005[1]. Shu-Hsien Laio, Hsu-hui Ho, and Hui Wen Lin, also succeeded in creating clusters of stocks in the Taiwanese market using financial returns [2]. The main drawback of both of these methods is that the clusters are based on results rather than inputs. Both methodologies rely on matching returns. Investors are still left to wonder why stocks are clustered rather than how, and extending the analysis to a new piece of information is impossible without an extensive history of returns and data. By using text from the documents themselves, important words and phrases are identified, and the analysis also applies to new articles and headlines. Effectively, users would be able to have a richer experience since the effects of these articles on returns could then be measured rather than clustering solely on returns.

III. Dataset and Features

The SEC provides a repository of all graphs, figures, and text from the 10-K and 10-Q reports submitted by all companies[3]. This data is provided on a quarterly basis, and includes such information of date of submission, type of document submitted, section of submission, text length, and finally the text. All text from the document including footnotes, headers, and chart titles. Each document was broken up by paragraph, and miscellaneous text snippets were also treated to their own row. For this reason, many rows included short pieces of text that were irrelevant and nonsensical. These fragments were often less than 40 characters long and included special characters and numbers from a chart, or serial date and filing number. By dropping these fragments, all the text backing one ticker was reconstituted as one large document. For this project the data from Q1 and Q2 2017 was used as the

training set, and Q3 2017 was used as the test set. While the dataset included all companies reporting to the SEC, only firms within the S&P 500 were considered for the project.

After reconstituting the 45,203 rows as one large document, the text had to be prepared for analysis. The first step in processing the text was in removing all of the numbers and special characters. The numbers while important in analyzing business performance will not translate well when tasked with finding correlations across companies due to their specificity. For LDA, a matrix of 9189 word counts had to be created. All stop words were dropped, as well as words that occurred in over a certain percentage of the documents. The intuition being that words that appeared in nearly every document likely had low explanatory power. The percentage was tuned to discover if there were differing impacts on performance, and this will be discussed further in the results section.

For NMF and K-Means clustering, the data had to be processed slightly different. For these two algorithms, numbers, special characters, high document frequency words, and stop words were removed just as in LDA processing. However, the final matrix consisted of TF-IDF rather than word counts. TF-IDF for each word was calculated using the following formula[citation]:

$$\text{TF-IDF} = \text{TF}(t,d) * \text{IDF}(t) \quad (1)$$

$$\text{IDF}(t) = \log[(1 + n_d)/(1 + \text{df}(t,d))] + 1 \quad (2)$$

TF(t,d) refers to the number of times a particular term appears in a document. IDF employs La Place smoothing which is why a 1 is added to both the numerator and denominator. N_d refers to the number of documents, and df(t,d) refers to the number of documents that the term appears in. The TF-IDF vectors were then normalized by the Euclidean norm. The resulting matrix was then ready for the NMF and K-means algorithm. For both methods only n-grams of length 1 were used.

IV. Models

For this study, three unsupervised learning models were used—K-Means clustering, NMF, and LDA. These models were picked for their relatively different approaches to unsupervised clustering.

A. K-Means Clustering [4]

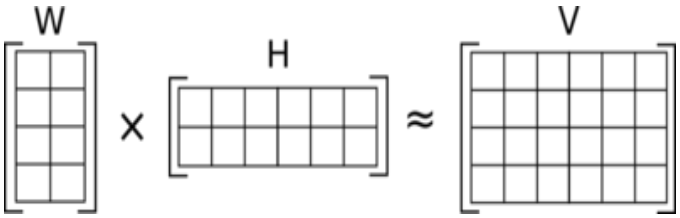
K-Means clustering was a good baseline model of prediction since it looked solely at the distances between documents and tried to place firms close together. For inputs it used a TF-IDF matrix, and used Euclidean distance as the distance metric.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

The above equation was used as the cost function, and the initialization of the centroids was random, but the number of centroids was inputted.

B. Non-negative Matrix Factorization (NMF)

NMF can be viewed as a dimensionality reduction algorithm.



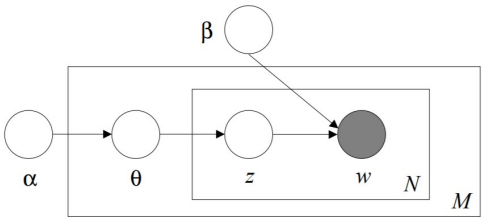
The input matrix V is approximated through the multiplication of the matrices W and H. Matrix W contains the principal vectors of importance, and Matrix H is the fractional combinations of the original data[5]. The columns of matrix W can be thought of as including the latent factors describing the interaction between the input variables. Therefore, from a clustering perspective, the columns of matrix W can be viewed as the topic centroids.

$$\min_{W \geq 0, H \geq 0} \|A - WH\|_F^2.$$

The above error function was used to find the matrices W and H using a Frobenius norm. The matrices W and H were minimized alternatively using an application of the EM algorithm[6]. Also a key aspect of NMF is that every matrix entry is non-negative.

C. Latent Dirichlet Allocation (LDA)[8]

LDA can be thought of as a model that assumes that each document is made up of a mixture of topics and each topic has its own distribution of word probabilities [7]. The following figure illustrates the idea:



With this figure [7] α represents a vector of values that parametrizes a Dirichlet distribution. As a hyperparameter the higher the value of α, the fewer number of topics each document contains. Conversely, a lower value means each document is composed of a larger mix of topics. The hyperparameter β works the same way but on word probabilities for any given topic. Z is the latent variable denoting the latent topics.

Given the hyperparameters α and β, the joint distribution of a topic mixture θ for any document is given by the following equation [7]:

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta)$$

V. Results

In order the determine the quality of the clusters created, the Rand Index was used as a metric of coherence. The companies were each labelled according to the sector that they are currently assigned to, and from there the quality of the clusters was calculated. There is an underlying assumption that the current sector definitions are accurate, and we will revisit this assumption.

The Rand Index is the odds that any randomly selected pair of items will match between two sets of clusters. Here one set of clusters is created by the model, and the second is the set of sectors used by the market. More formally defined:

$$R = A+B / (A+B+C+D)$$

A represents the number of pairs of elements that are in the same subset in both sets of clusters. B Represents the number of pairs of elements that are in different subsets in both clusters. C represents the number of pairs of elements that are in the same subset in one set of clusters, and in different subsets in the second set of clusters. D represents the number of pairs of elements that are in different subsets in the first set of clusters, but are in the same subset in the second set of clusters.

While there are 11 official market sectors, there are 24 more subsectors[9]. Some of these breaks are quite natural as heavy machinery companies and airlines both fall in the same category – Industrials. The same can be said of the broadcasting companies and Google, Akamai, and Facebook all placed under Information Technology. Therefore, the number of topics was varied to determine how many sectors the model assumed that there were.

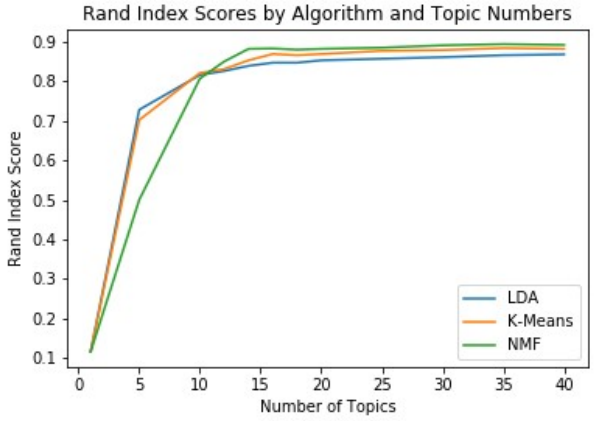
TABLE I. Document Frequency of 10%

Model	Rand Index Scores by Topic Number and Model						
	5	10	12	14	16	18	20
K-Mean	0.702	0.821	0.831	0.853	0.869	0.866	0.869
NMF	0.500	0.807	0.850	0.882	0.883	0.880	0.882
LDA	0.728	0.816	0.826	0.839	0.847	0.847	0.853

A. Analysis

As a ground truth, if all the companies were place into one large cluster, the Rand Index across all three models would be 0.116. Clearly, we can see that all of the models are a significant upgrade. One of the more surprising results was that NMF was the best performer. I had expected LDA to outperform given the additional freedom in parameters it had, and its ability to find topic mixtures.

All three models seemed to plateau between the 14-18 topic range. This implies that the models believe that there are more than 11 market sectors present. Given the diversity of some of the sectors this does make sense. All three models plateau in that range and there is very little increase in the Rand Index beyond that point. This chart further illustrates that point as it goes out to 40 topics.



The terms used were limited by document frequency. The intuition is that if a term is present in the majority of documents then it likely has very little explanatory power. Restricting and relaxing the limit to document frequency did have minor impacts on the results. A maximum of 5% document frequency led to a slower plateau, and a plateau that was lower than at 10% and 20% document frequency when looking at topics=16 which seemed to be the rough area where the 10% document frequency plateaued.

TABLE II. Document Frequency of 5%

Model	Rand Index Scores by Topic Number and Model						
	5	10	12	14	16	18	20
K-Mean	0.730	0.816	0.838	0.868	0.855	0.868	0.860
NMF	0.479	0.674	0.825	0.838	0.852	0.876	0.884
LDA	0.722	0.809	0.820	0.831	0.838	0.845	0.846

TABLE III. Document Frequency of 20%

Model	Rand Index Scores by Topic Number and Model						
	5	10	12	14	16	18	20
K-Mean	0.727	0.828	0.833	0.849	0.868	0.866	0.871
NMF	0.641	0.823	0.846	0.870	0.881	0.872	0.893
LDA	0.730	0.813	0.819	0.841	0.849	0.835	0.854

NMF in particular takes longer to plateau given an increase in topics, and this makes sense given that it has less information to work with to create matrix vectors. Effectively we are compressing the same information twice. K-Means is the most stable across document frequencies. Overall while there is a negative impact it is not as large as I would've expected and shows that specific low frequency words have the majority of the explanatory power. Still surprising is that NMF is the best performer with LDA being the worst. While they are

close, I would have expected LDA to outperform. One possible reason for this is a mistuned hyperparameter alpha or beta. The default parameter setting for alpha and beta is 1/number of topics[8]. Tuning the alphas and betas resulted in the following:

TABLE IV. DOCUMENT FREQUENCY OF 10% AND 16 TOPICS

Model	Rand Index Scores by Alphas and Beta						
	0.005	**0.01**	**0.03**	**0.06**	**0.1**	**0.15**	**0.2**
Alpha	0.843	0.843	0.841	0.843	0.843	0.839	0.844
Beta	0.839	0.842	0.841	0.837	0.839	0.842	0.844

Modulating the alpha and betas had a very minimal impact on the Rand Index. There is an assumption that actual topics are modelled by a symmetric Dirichlet which may be incorrect. Allowing for asymmetric Dirichlet in relation to the alpha and beta could result in stronger performances for the LDA model. Especially considering that the sector distribution is not equal—some sectors have twice as many companies as others. This is something that requires further research.

An alternative explanation for the underperformance of the LDA model in comparison to the NMF and K-Means models may be in the structure of the reports themselves. LDA is the most adaptive of the models, and may be picking up something else in the reports that is not perfectly correlated to company sector. Company performance is a large part of these reports. As a result, a struggling consumer retail company may be addressing a lot of the same factors that a struggling energy company is addressing. They both may spend a large part of their report speaking about debt covenants, loan rates, and cash reserves. While these are important to all businesses the term frequencies will likely be higher in struggling companies, thus grouping them together. This business cycle component of the reports may be a confounding factor for all three models, with LDA impacted the most given its flexibility.

The Rand Index scores are all based on how closely the clusters generated by the models matched existing market sectors. Given that all three models clearly outperformed random chance, there is subjectivity in which model is actually the best. The highest Rand Index score may quantitatively suggest one model to be superior to another, but there is still a large qualitative component to this problem. Not all of the current sectors necessarily make complete sense. Broadcast companies and Telecommunications companies are separated in different sectors while they could be thought of as very close in nature. Perhaps, closer than several internet companies that the Broadcast firms are paired with. To further evaluate the clusters generated, each cluster and company assigned to it would have to be subjectively evaluated not only on what the company did but also the part of the business cycle it was in.

On the test set which comprised all 14,862 documents from Quarter 3 of 2017 that were longer than 40 characters and pertained to S&P 500 companies, the 16-topic, 10% document word frequency version of LDA, NMF, and K-Means were chosen. The hyperparameters on LDA were left at default values.

TABLE V. TEST VERSUS TRAIN, TOPICS=16, AND DOCUMENT FREQUENCY= 0.1

Model	Rand Index Scores for Test and Train		
	LDA	**NMF**	**K-Means**
Test	0.846	0.882	0.861
Train	0.847	0.883	0.869

There was very good generalization from the train to the test set for all 3 models. There is very little evidence of overfitting, and LDA and NMF generalize very well. NMF is still the top performer quantitatively, but more work is necessary to qualitatively select the best model.

VI. CONCLUSION

NLP combined with unsupervised learning methods is successful in creating market sectors relatively similar to existing ones. The quality of the sectors created by the models is highly subjective, since even the current sectors are open to some interpretation. With a cursory overview many of the pharmaceutical companies were grouped together as were the financial companies, and energy companies. Most of the sectors generated made sense. Quantitatively a Rand Index score over 0.84 for all 3 models indicates good clustering. This study as a proof of concept was a success.

Some simple extensions to this product with additional resources would be to cluster based on subsectors as well to see if the additional granularity results in a better fit. Looking at each generated cluster closely to see if it makes sense and why the companies were placed together. There is no reason that the current market sectors are to be seen as the ideal although they are a convenient baseline. Incorporating a larger amount of data may also help. This study was limited to 2017 largely to ensure that it could run on my personal laptop, so adding more data may help, and may reduce any business cycle effects. Incorporating an asymmetric Dirichlet prior for alphas is also something worth pursuing since the official sectors are not evenly distributed. Trying out n-grams larger than 1 in the future would also be interesting.

As a proof of concept this study proves that clusters can be created from the 10-K and 10-Q reports filed to the SEC. However, this study is a stepping stone for further research. Ultimately, I would be like to be able to take news articles and headlines and identify which companies are likely to be most affected by them and by how much. The first component of finding clusters of companies has been done. Identifying tone within newspaper articles is the next step, and then mapping them to the clusters of companies. In mapping articles to companies we may find that increasing the number of topics helps since it means that there are smaller groups of more similar companies and each article may have more explanatory power.

VII. REFERENCES

[1] Kein A.J. Doherty, Rod G. Adams, Neil Davey, and Wanida Pensuwon, "Hierarchical Topological Clustering Learns Stock Market Sectors," Computational Intelligence Methods and Applications, 2005 ISCS Conference on, December 15 2006.

[2] Shu-Hsien Liao, Hsu-hui Ho, and Hui-wen Lin, "Mining Stock Category Association and Cluster on Taiwan Stock Market," Expert Systems with Applications 35, 2008, pp.19-29.

[3] "Financial Statement and Notes Data Sets." *Financial Statement and Notes Dataset*, 4 Oct. 2017, www.sec.gov/dera/data/financial-statement-and-notes-data-set.html.

[4] "Sklearn.cluster.KMeans." *Sklearn.cluster.KMeans scikit-Learn 0.19.1 documentation*,2017,scikitlearn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[5] Lopes N., Ribeiro B. (2015) Non-Negative Matrix Factorization (NMF). In: Machine Learning for Adaptive Many-Core Machines - A Practical Approach. Studies in Big Data, vol 7. Springer, Cham

[6] "Sklearn.decomposition.NMF." *Sklearn.decomposition.NMF scikit-Learn0.19.1documentation*,2017,scikitlearn.org/stable/modules/generated/sklearn.decomposition.NMF.html.

[7] David M. Blei, Andrew Y. Ng, Michael I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research 3, 2003, pp.993-1022

[8] "Gensim: topic modelling for humans." *Radim Rehurek: Machine learning consulting*, 17 Dec. 2017, radimrehurek.com/gensim/models/ldamodel.html.

[9] "GICS." *MSCI*, 2017, www.msci.com/gics.