Predicting Stock Returns From Stock Forums

Sunny Yang, Wiley Wu, Evan Wang, Jasper Wang

May 9, 2024

Abstract

The explosive growth of retail investors in the stock market, which was primarily influenced by social media platforms, raises questions about the predictive power of social media on market dynamics. Our project explores whether or not posts on popular stock-related forums, such as Reddit's r/Stocks and r/WallStreetBets, can predict stock market returns. Our project has adapted the Latent Dirichlet Allocation (LDA) and supervised LDA models to the unique corpus of social media posts (specifically stock forums on Reddit), building on the foundational methods initially applied in the study 'Choosing News Topics to Explain Stock Market Returns' [6] by Glasserman and colleagues. We have then applied a logistic regression model to forecast stock price movements based off the output from our fitted sLDA models. Additionally, we employed Bidirectional Encoder Representations from Transformers (BERT) for sentiment analysis, providing a baseline for comparison with our sLDA results, which have demonstrated moderate performance so far. Furthermore, we explored how similar these stock-posting forums are in order to better understand how posts from different sources can affect the accuracy of our model. Posts varied wildly in tone and format, which made comparing predictions from more formal subreddits such as r/Stocks with less formal ones like r/WallStreetBets very interesting. By comparing our approach with previous studies, we are able to determine the possible utility of using social media topics to forecast stock market trends, which can contribute to the development of more advanced quantitative trading algorithms that account for the increasing influence of retail investors.

1 Introduction

Since 2020, the landscape of the stock market has witnessed a pronounced influx of retail investors, contributing to as much as twenty-five percent of total options trading activity by 2022 [8]. This shift is notably propelled by the increasing use of social media platforms, such as Reddit, where retail investors exchange insights and strategies concerning stock movements. The burgeoning popularity of these platforms is evidenced by the explosive growth of communities like r/WallStreetBets and r/Stocks, which saw their user bases expand to nearly eight million and two million respectively in 2021 [5] [4].

Concurrently, the rise of mobile-centric trading platforms, notably Robinhood, has paralleled this trend. Robinhood itself reported an increase of ten million users in the same period, reaching a zenith during the meme stock phenomenon of January 2021 [10]. A pivotal moment in this narrative was the Gamestop (GME) saga, where the concerted actions of r/WallStreetBets members notably led to a dramatic surge in GME's stock price. This event significantly impacted major hedge funds, including Melvin Capital and D1 Capital Partners, which faced substantial losses exceeding six billion dollars due to their short positions [3] [7].

The repercussions of these market dynamics extended beyond financial losses, prompting a legislative and regulatory review when Robinhood controversially restricted trading on stocks like GME, AMC, and BlackBerry during critical trading periods in late January 2021

[9]. These actions, driven by collective retail investor behavior on platforms such as r/WallStreetBets, underscore the profound impact of social media on market movements.

Given this context, our research aims to ascertain whether Reddit posts contain predictive signals that could anticipate such market movements. This study represents an initial endeavor to develop a predictive model leveraging the rich textual data from social media, specifically analyzing how these narratives influence stock prices. Understanding these dynamics is crucial, as needing to model retail investor behaviors is becoming increasingly integral to quantitative trading strategies. Moreover, this investigation will explore which metrics are deemed most critical by different investor segments, aiding in the development of more tailored and effective trading algorithms. As the global equity market becomes more inclusive of retail investors, it is imperative to comprehend how their perceptions and media interactions influence market valuations and the strategic development of financial algorithms.

2 Problem Setting

Topic models are probabilistic algorithms designed to uncover hidden thematic structures within large collections of documents. They help interpret news sentiment by identifying prevalent topics, enabling researchers to analyze relationships between specific topics and market trends. Considerable research has been conducted on leveraging news and social media sentiment to predict market trends. Some recent studies have explored the correlation between public sentiment and market movements, revealing that significant shifts in sentiment often precede fluctuations in stock prices. Professor Paul Glasserman's research paper, *Choosing News Topics to Explain Stock Market Returns*, examines the relationship between specific news topics and market returns [2]. Glasserman specifically identifies relevant news topics that are closely linked to stock performance, highlighting their potential predictive power. However, there has been limited research on using social media post topics to predict a company's stock return on the release date of the post. Applying topic models to forecast market return fluctuations is an attempt to deepen the relationship between social media post topics and the corresponding post content. We primarily utilize two models, which are described in the next section.

2.1 Model Selection

Numerous topic models are available in the field of text analysis. This research begins with a theoretical evaluation of several prominent topic models. Our analysis includes a variety of models: the Unigram Model, which treats each word in a document independently; the Mixture of Unigrams Model, where each document is assumed to be a mixture of topics; Probabilistic Latent Semantic Indexing (pLSI) also known as the Aspect Model, which also views documents as mixtures of topics but incorporates a probabilistic approach; the Latent Dirichlet Allocation (LDA) Model, which allows for more complex distributions of topics within documents; and the Supervised LDA (SLDA) Model, which extends the LDA framework to include supervision for predictive tasks. This phase of the research aims to determine which of these models is most suitable for effectively handling and analyzing our specific dataset.

A few definitions: We define a corpus as a collection of social media posts, with each document consisting of a series of words. Mathematically, the words within a document are denoted as $W_{1:n}$. Given a total of j documents, we define a large text matrix as $W_{i,j}$. Words are drawn from a dictionary of size V.

2.1.1 Unigram Model

The Unigram Model relies on the assumption that words are drawn independently from a multinomial distribution. This model considers that all text is generated from a single distribution. The graphical representation of this model is shown in Appendix A. The probability distribution is given by:

$$p(\mathbf{w}) = \sum_{n=1}^{N} p(w_n)$$

2.1.2 Mixture of Unigrams Models

Expanding on the Unigram Model, the Mixture of Unigrams Model introduces a discrete random variable[2] z. This model allows each document the freedom to choose from a single topic z. The probability distribution is as follows:

$$p(\mathbf{w}) = \sum_{z} p(z) \sum_{n=1}^{N} p(w_n|z)$$

2.1.3 pLSI/Aspect Models

The pLSI model[2] is a more complex latent factor model. It introduces a labeled random variable d, such that the label d and words w_n are conditionally independent given an unobserved random variable z. The probability distribution is of the following form:

$$p(d, w_n) = \sum_{z} p(d) \sum_{n=1}^{N} p(w_n|z)p(z|d)$$

Although the model fits the topic modeling task well, it has limitations; the random variable d is equivalent to the number of documents in the training set. This model is unable to naturally fit into any text document that does not inherently have a label.

2.1.4 Plain LDA model

Given that the Latent Dirichlet Allocation (LDA) model[2] serves as the baseline model in our research, it is essential to detail this model in our definitions section. Here are some fundamental assumptions we have made: Firstly, we assume $N \sim \text{Poisson}(\xi)$, implying that the number of words appearing in a Reddit post follows a Poisson distribution. This distribution assumption is considered trivial and does not significantly influence our model estimation. Secondly, we select $\theta \sim \text{Dirichlet}(\alpha)$. The advantage of choosing a Dirichlet distribution is its flexibility in encoding various topic combinations. More importantly, it promotes competing sparsity over a probability vector of size k, suggesting that a document is more likely to originate predominantly from a few major topics. This will be illustrated in the graphical representation in Section 3. Finally, for each of the N topics w_n, we choose a topic $z_n \sim \text{Multinomial}(\theta)$, and subsequently, a word w_n from $p(t_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n. Based on the graphical model, we could also write down the probability distribution:

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

$$p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta$$

The graphical representation of LDA model:

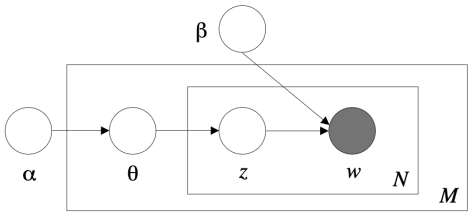

Figure 1: LDA model

2.1.5 sLDA Model

In sLDA[1], each social media post is assumed to be generated from a mixture of topics, where each topic is a distribution over words. The social media post's topics are represented by a vector of topic proportions θ, which sum to 1. These proportions are typically drawn from a Dirichlet distribution parameterized by α, i.e., $\theta \sim \text{Dir}(\alpha)$.

The generative model for the words in sLDA is similar to that of unsupervised LDA. However, sLDA extends LDA by also modeling a response variable y associated with each social media post. The response variable is typically modeled as a function of the topic proportions:

$$y \sim f(\theta, \beta, \eta, \delta)$$

$$p(\theta, z, w \mid \eta, \delta, \alpha, \beta) = p(\theta \mid \alpha)p(y \mid \eta, \delta)\sum_{n=1}^{N} p(z_n \mid \theta)p(w_n \mid z_n, \beta)$$

The graphical representation of sLDA model:

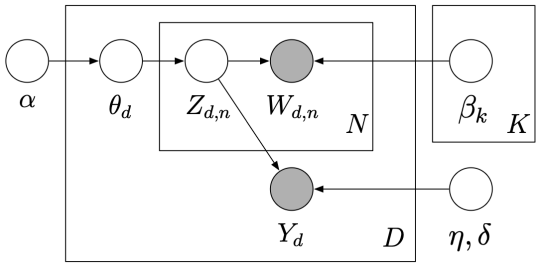

Figure 2: sLDA model

In the formulation of the supervised Latent Dirichlet Allocation (sLDA) model, several parameters define the relationship between the model components and the response variable. The parameter θ denotes the topic proportions within a social media post and β corresponds to the parameters governing the word distributions across different topics. η represents a parameter vector that links the topics and the response variable y. Lastly, δ serves as a dispersion parameter, influencing the variability of the response variable based on its underlying distribution.

In our research, we focused on predicting market trends, specifically whether they will rise or fall. Therefore, we tailored a logistic regression model to fit the conditional distribution. The generalized linear regression model offers the versatility to handle various regression tasks, allowing us to accommodate different types of response variables in future analyses.

2.2 A Comprehension of Different Topic Model

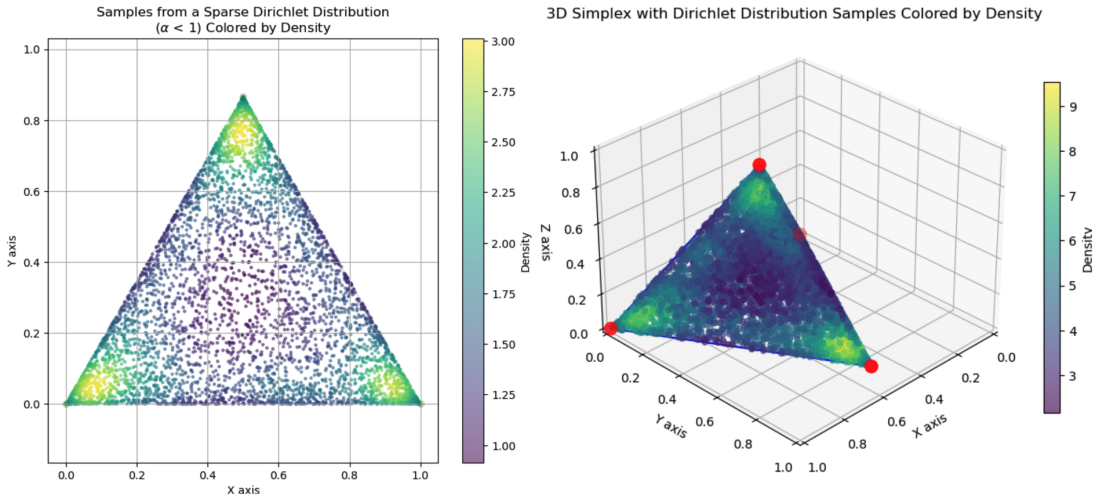

Figure 3: Using Simplex to demonstrate the topic modeling novice case

When exploring the geometric aspects of latent topic models, it's pivotal to comprehend how documents are depicted within these frameworks. The discussed models—unigram, mixture of unigrams, pLSI, and LDA - all function within the realm of topic distribution spaces, which are geometrically mapped onto a $v-1$ simplex, The 2-d simplex on the right represent a three topic case, and the 3-d simplex on the right represent a three topic or 2D simplex case, and four topic 3D simplex case. Models like LDA exhibit the capability to encompass the entire scope of the simplex space, offering a versatile representation of documents through a probabilistic blend of topics. This flexibility is not typically afforded by other models, which may adhere more strictly to predefined topic distributions. Moreover, the inherent sparsity observed in these models enhances their interpret ability, as it allows for distinct topic delineation, thereby aiding in clearer understanding and analysis of the document topics.

3 Methodology

3.1 Data Collection

We sourced our data directly from Reddit stock posting subreddits, specifically posts from r/Stocks, r/StockMarket, r/WallStreetBets, and r/Investing, since these are some of the most active and well-known stock subreddits. Our dataset was comprised of 4000 posts; we scraped 1000 posts from each of the forums using a Reddit bot that grabbed all relevant information: post author, the body of the post (also referred to as 'selftext'), the post's ID, and the date of the post's submission. When collecting posts, we focused on older posts that were tagged with a flair of "Company Analysis" (or the equivalent). Our rationale behind this is twofold: first, posts with these flairs (or tags) generally offer both a more serious sentiment and more complex analysis, which helps us eliminate a majority of low-quality submissions (i.e. posts with only a few words, or posts with just a single image and a caption). Secondly, by using older posts, we were able to set a three-month interval (one financial quarter) as a baseline for our stock price predictions.

3.2 Pre-Processing

After gathering the data and sorting it by forum (subreddit), we then began processing the corpus using the bag-of-words approach. This approach allows us to assume that all words are weighted equally and order doesn't matter, which allows us to drastically reduce computational efforts and generally lose little meaning. After tokenizing all the posts, we then use a pre-built NLTK dictionary, augmented with common financial terms, to remove stop words. This augmentation allows us to avoid filtering out words that are important in a financial context, which is crucial given that we are analyzing stock market forums. Additionally, we have also fine-tuned the amount of punctuation that was removed, in order to maintain meaningful information (i.e. numbers without decimals can become functionally meaningless).

3.3 Data Labeling

Since we primarily use sLDA as our main model, labeling the data accurately is crucial. There are two key labeling tasks involved. First, we need to determine and label whether the stance of each post is "long" or "short" in market trading terms. Additionally, we must label the changes in market prices over the period discussed in the posts, and if missing, we'll use 1 quarter (3 months) ahead. Labels generated through this process are considered the definitive labels for the positions' stances.

3.4 Sentiment Analysis

After labeling our data, we then used a Bidirectional Encoder Representations from Transformers (BERT) model for sentiment analysis. BERT models are commonly used as a baseline in NLP because of their ability to capture contextual word representations through bidirectional training, which improves text understanding. We have also chosen to use BERT as our baseline sentiment analysis model. After training BERT on a pre-labeled financial dataset, we then performed sentiment analysis for each post, labeling each post either positive or negative. If a post is labeled as positive, then we interpret this as predicting the post will have a long position (the post expects a stock's value to rise in the future). If a post is labeled as negative, then we interpret this as predicting the post will have a short position (the post expects a stock's value to fall in the future).

3.5 Similarity Metrics

We were also interested in how similar or diverse the different subreddits are. We used several diversity, complexity, and similarity metrics to evaluate and compare the different forums: type-token ratio (TTR), average Flesch Reading Ease (FRE) score, Measure of Textual LD (MTLD), TF-IDF tables, and cosine similarity (all posts aggregated by forum). Type-Token Ratio (TTR) is a measure of lexical diversity in a text, calculated by dividing the number of unique words (types) by the total number of words (tokens). A higher TTR indicates greater lexical diversity, while a lower TTR suggests repetition or a limited vocabulary. The average Flesch Reading Ease (FRE) score is a measure of the readability of a text, calculated by assessing the average sentence length and the average number of syllables per word. A higher score indicates that the text is easier to read. MTLD (Measure of Textual Lexical Diversity) is a measure that assesses the lexical diversity of a text by calculating the mean length of sequential word strings that maintain a given Type-Token Ratio (TTR) threshold. MTLD is less affected by text length, making it a more reliable indicator of vocabulary variety. TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic used to evaluate the

importance of a word in a document relative to a collection of documents. It is calculated by multiplying the term frequency (TF) of a word in a document by the inverse document frequency (IDF), which measures how rare the word is across the entire document set. Cosine similarity is a metric used to measure the similarity between two vectors by calculating the cosine of the angle between them. It ranges from -1 to 1, where a value closer to 1 indicates high similarity, and values closer to -1 signify strong dissimilarity. Figures for the similarity metrics are included in the appendix.

3.6 First LDA Layer: Predicting Post Position

Our project utilizes an sLDA model in two layers: first, we train the LDA model on the dataset, and train a logistic regression model to predict the sentiment of the model (to compare against our baseline BERT sentiment analysis). The output is either 0 or 1; an output of 1 means that the model predicts that the post has a long position; an output of 0 means that the model predicts that the post has a short position. We then compare our predictions against the post's actual position.

3.7 Second LDA Layer: Predicting Stock Price Movement

Then, we use the LDA model's output and labeled stock price movements to train a logistic regression model to predict the stock price movement, based on the text data (separate from the post's position). The output is either 0 or 1; an output of 1 means that the model predicts that the stock price will increase; an output of 0 means that the model predicts that the stock price will decrease. We then compare our predictions against the actual stock price movement a financial quarter after the post's submission.

4 Model Evaluation

Our current results show that our sLDA model has an accuracy of 79.16% (shown in Figure 4) in correctly determining the actual post's position, in comparison to our BERT sentiment analysis, which has an accuracy of 67.2% in correctly determining the actual post's position, which is a notable 11.89% improvement in accuracy. The current results show that our sLDA model has a 70.83% accuracy (shown in Figure 5) in correctly predicting the stock price direction for a three-month period.

```
Model score: 0.7916666666666666
Topic 0: right just long earnings don sales time market auto like year going revenue stock business
Topic 1: homedepot sales imo wallstreetbets disney 5g just google microsoft nbsp x200b going cloud
Topic 2: virus price time market ve buy money dd amp calls good don earnings going people like just
```

Figure 4: Model accuracy & Important Topics for Position (Sentiment) Prediction.

```
Model score: 0.7083333333333334
Topic 0: options 10 right year time don massive supply market 2020 ve earnings debt com
Topic 1: price market long don business year revenue money png stock going time earnings
Topic 2: amd news cloud earnings post calls 2020 azure going just good dd wallstreetbets
```

Figure 5: Model accuracy & Important Topics for Stock Price Prediction.

Additionally, our analysis of the forums has shown that the most lexically diverse forum in terms of average MTLD score is r/WallstreetBets (197.81), while the most complex forum

in terms of average FRE score is r/Stocks (-183.42). r/Investing has the highest TTR (0.742) but the lowest average tokens, types, and average FRE score (-65.09). The TF-IDF tables show extremely similar results for the most important terms for each forum, indicating a high degree of similarity between the posts. r/Wallstreetbets and r/Stocks both place a high emphasis on "earnings", where both r/Investing and r/StockMarket do not.

5 Discussion

However, our model is heavily limited by the limitations of our data; given the time constraints, our dataset is comparatively miniscule in relation to the corpus Professor Glasserman used in his study. This is partly due to the fact that collecting and labeling this data is very time-consuming and also partly due to the fact that we simply do not have enough resources to handle computation on that level; a corpus of 90,000 articles (or in our case, posts) can generate enormously large datasets, and our sLDA models would take an incredibly long time to fit and transform our data.

Furthermore, another limitation of our current model is that we only had a very limited time period in which we collected data from. Given that the retail investor phenomenon has happened only since 2020, we have few years of both financial data, and data from retail investors online. Additionally, another factor to consider is the general market trend that occurred during this time period: the pandemic "bubble" may have caused our model to have an outlier performance as a result of the general market dynamics, instead of our data. Additionally, much of finance relies on time-series data, making it significantly more difficult to obtain independent and uncorrelated data points especially as we begin to scale up.

6 Conclusion

In conclusion, our project has demonstrated the efficacy of social media data as an influential source for quantitative trading algorithms. Our results show a substantial improvement in prediction accuracy compared to traditional sentiment analysis methods. However, our project's limitations are especially apparent when compared to Prof. Glasserman's study and results. Notably, we need to further refine our preprocessing procedures (i.e. stopwords, fine-tune BERT) and address our data limitations. Additionally, our study differs in source and outputs, but are comparatively similar in accuracy (roughly 75%). Our project analyzes user-generated content from Reddit, which is much more informal and highly reactive. In contrast, Prof. Glasserman's study uses structured news articles from more traditional sources (i.e. magazines, newspapers). Ultimately, our project is primarily focused on implementing models that influence trading strategies, while Prof. Glasserman's work analyzes how news shapes market perceptions and outcomes.

References

[1] David M. Blei and Jon D. McAuliffe. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2007.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] Juliet Chung. *WSJ News Exclusive — Melvin Capital Lost 53*

[4] Frontpagemetrics.com. /R/Stocks Metrics (Stocks)*, 2021.*

[5] Frontpagemetrics.com. /R/Wallstreetbets Metrics (Wallstreetbets)*, 2021.*

[6] Paul Glasserman, Kriste Krstovski, Harry Mamaysky, and Paul Laliberte. Choosing news topics to explain stock market returns. Proceedings of the ACM International Conference on AI in Finance (ICAIF-2020)*, 1, October 2020.*

[7] Burton Karsh et al. Dan Sundheim's $20 Billion D1 Capital Loses about 20

[8] Yun Li. Options Trading Activity Hits Record Powered by Retail Investors, but Most Are Playing a Losing Game*, 2021.*

[9] Chris Mills Rodrigo. Robinhood Restricts Trading of Companies Targeted by Reddit Users*, 2021.*

[10] statista.com. Number of Robinhood Active Monthly Users with Average Revenue Per Users (ARPU) from 2014 to 2023*, 2024.*

7 Appendix A

The graphical model of Topic Models

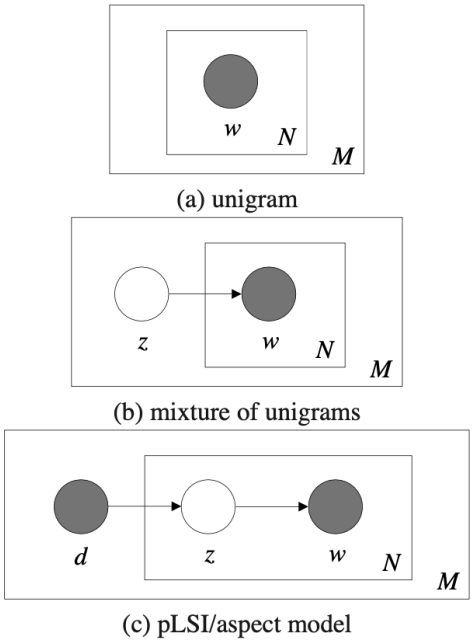

Figure 6: Unigram, Mixture of Unigram, and PLS Models.

8 Appendix B

	TTR	Median Tokens	Average Tokens	Average Types	Average FRE Score	Average MTLD Score
Investing	0.742489	79.0	117.295238	87.090476	-65.094929	138.996392
Stock Market	0.679691	141.5	207.809859	141.246479	-166.509789	140.451113
Stocks	0.644805	96.0	232.202864	149.725537	-183.425561	139.546256
WSB	0.716130	139.0	224.641913	160.872838	-177.837294	197.818914

Figure 7: Table with TTR, average FRE score, and MTLD.

	Investing		Stock Market		Stocks		WSB
stock	0.248836	price	0.385012	stock	0.435031	stock	0.207319
fund	0.200345	stock	0.293545	price	0.163986	going	0.172034
rate	0.178651	share	0.146773	week	0.152018	people	0.153632
money	0.168443	level	0.144645	share	0.151371	price	0.144684
share	0.144197	support	0.124438	earnings	0.147813	time	0.140464
price	0.143559	option	0.124438	time	0.142638	call	0.138269
bond	0.139731	major	0.123374	rate	0.138434	go	0.132191
time	0.136541	rate	0.119120	down	0.107060	earnings	0.124425
account	0.119952	high	0.105293	today	0.100591	make	0.122737
investment	0.116761	may	0.101039	trading	0.099620	share	0.122061

Figure 8: TF-IDF Tables for all 4 forums.

	Investing	Stock Market	Stocks	WSB
Investing	1.000000	0.718359	0.813137	0.746786
Stock Market	0.718359	1.000000	0.796118	0.696306
Stocks	0.813137	0.796118	1.000000	0.776557
WSB	0.746786	0.696306	0.776557	1.000000

Figure 9: Cosine Similarity between all 4 forums.