Derick Ariyam
Professor Karno
ENG 590
October 23, 2012

A New Theoretical Approach to Measuring Subjective Value in the Stock Market

Friday, May 18, 2012 was a historical day for the stock market and the social networking behemoth Facebook. On that day, Facebook officially became a public company. For the price of 38$ (USD) one could own a small piece of the corporation (or, approximately 1/421,000,000th of a piece). The initial offering price established by underwriters set the total market capitalization (or perceived net value) of this company at 104 billion dollars. However, in a few short weeks of exposure to the market that initial share price of 38$ changed dramatically. Facebook stock began to decline—rapidly. In three months, the same share of Facebook was now valued at only 19$. The company effectively shed half of its original value (or 52 billion dollars).

As media and analysts began to investigate what went wrong with the stock, it was clear that no material change in Facebook could account for such a great loss of value. The company's quantifiable properties, or what investors call "the fundamentals" (assets, debt, cash, investments etc.), were not much different than they were at the initial public offering (IPO) date. What *was* different was that more collective minds now had a stake in the company. The sentiments of millions of individual, corporate, and institutional subjectivities were now free to impose their impressions on the stock.

The original IPO value of 38$ set by underwriters was based on fundamentals and future growth potential. But underwriters had no means to accurately predict overall market sentiment. There are currently no models available that can simulate sentiment (or at least not accurately).The Facebook IPO disaster was a reminder that this mercurial, unquantifiable, elusive metric forged by emotive responses and subjectivity holds such an incredible bearing on the stock market. It is both interesting and unsettling that this force can have such a significant impact while remaining difficult—if not impossible—to quantify.

However, despite this difficulty, new efforts are being made to try and measure the subjective influences on the market through the aid of computational algorithms. This work is fairly nascent in its development and has been brought about primarily by the wide adoption of social networking websites like Facebook and Twitter as well as the availability of real-time news and headline feeds over the web.

Social networking platforms collect massive amounts of data through the actions of their users. Facebook recently released statistics of just how much data is collected on an average day. The numbers are staggering. As of September 30th 2012, Facebook has over 1.01 billion users (Olanoff). The company estimates that users upload about 300 million photos and generate 2.5 billion data points (through status updates, wall posts) every day, amounting to an overall influx of 500 terabytes of new data (Kern).

Though not nearly as large as Facebook, Twitter also boasts a hefty following. The latest statistics by that company state that there are over 140 million users that use the service. The website generates an average of 350 million "tweets" a day. A "tweet" is a 140 character statement a user can make on the site, which is then recorded and broadcast to anyone following that user (Wasserman).

While the type of data Facebook and Twitter collects, such as wall posts, "likes," and tweets, may seem at first irrelevant or at least too quotidian for serious analysis, its sheer volume as well as the site's enormous user population allow for interesting data-mining potential. Both sites offer access to some of this data through publicly available "application programming interfaces" (APIs). These interfaces allows programmers to query the social networking databases directly and receive raw data sets that can be downloaded and analyzed. Through the use of APIs and developments in natural language processing techniques, researchers are beginning to use this data for a variety of interesting analysis, including as a means to calculate the elusive metric of subjective impact on the stock market.

Stock analysts have recently begun to use Twitter data to determine general consumer sentiment for specific companies. To do this, a programmer would first write code to interface with the Twitter API. They would then extract all tweets pertaining to a given stock symbol within a set of parameters (date, location, keyword). The data is then stored locally in a relational database, and each 140 character tweet is run through an algorithm to determine a sentiment score: usually positive, negative, or neutral. There are various heuristics available to calculate sentiment score—some more or less complex than others—but most involve some form of word counting. Generally, the algorithm places various sentiment "weights" to each word used in a given tweet. All conjunctions, articles, pronouns—or "noise words"— are removed from analysis and words and punctuation that may express strong emotions, like "love," "hate," emoticons and exclamations, are aggregated to determine the general sentiment of a tweet.

There are a variety ways this metric can be used to make stock decisions. Since Twitter data is real-time, in theory if one can gauge negative or positive sentiment from millions of tweets about a particular security, one can use that information to make immediate decisions on the stock market. Another approach is to monitor sentiment over time on a stock; if consumers appear to have overly negative or overly positive experiences of a company, one could use that information to predict earnings potential on a stock before the company's official quarterly or year-end statement. There are more complex techniques as well. One popular one involves options straddling, where an investor essentially bets that a stock will have a major price change. Using Twitter data, the observation of a disproportionate number of tweets of either a positive or negative nature will enable the investor to profit regardless of the direction of the price change, as long as that change was significant (Chakoumakos).

Mining language data for sentiment may seem like an exciting and straightforward prospect, but it's not without problems. The immediate challenge in analysis derives from the fact that this class of data—language—is non-empirical. Language can be subtle, ironic, misleading and sarcastic. Computers are not well-equipped to interpret the nuances of language. There are currently no reliable methods of interpreting this through computational algorithms. Even in the science fiction series *Star Trek: Next Generation*, where science has evolved to the state where matter can be transported through air and ships can travel at the speed of light, the most sophisticated computers, like the android Data, still struggle to accurately interpret basic irony and humor.

Additionally, this problem of analyzing sentiment in language through large data sets falls within the domain of two opposing academic camps: the sciences and the humanities. The interpretation of language is a skill practiced and studied in departments like English and Writing, while computational algorithms and statistical models are researched in Math and Computer Science departments. However, recent scholarship in the new emergent field of the Digital Humanities endeavors to produce linkages between these divergent academic interests. The goal is to provoke new and innovative ways to solve issues like sentiment analysis.

Some of the issues often studied and rehearsed in humanities departments may help illuminate issues with current implementations of sentiment analysis. The first issue is the reduction of sentiment to fixed categories, and/or binary logic: positive versus negative sentiment. There are obvious problems with such a drastic curtailment of the complexities of emotive response to finite categories. Something is obviously lost in this translation. However, we know from digitalization that translation from analog to binary digits, while reductive, the perception of loss can be mitigated by increasing the resolution of data. For example, by translating a sound wave which is inherently continuous into a discrete value effectively truncates the original data. But, when the resolution or "sample rate" is increased to a high enough level (typically 44,000 per/second) the loss is imperceptible. The same principle applies to digital photos and video. The question to consider is whether the same principles of digitalization can apply to language as well? If we increase the number of language data sets to a very high resolution, will that mitigate loss during the process of assigning sentiment? If so, at what resolution will this occur: how many tweets must one process to lower the significance of the reduction? Using historical stock data, one may be able to approximate what this value looks like by back-testing sentiment analysis across many years of stock data. However, while decades of historical data are available for stocks, the same is not true for Twitter or Facebook data feeds. It will take more years of accumulated data to draw connections.

Secondly, interpreting language is unsettled ground to begin with. The classic semiotic and linguistic theories developed by thinkers like Ferdinand de Saussure, Jacques Lacan and Jacques Derrida all point to the inherent indeterminacy of language. Saussure was among the first to suggest that the relationship between the signifier and signified is arbitrary: words being merely signs that point to psychological concepts which are inherently subjective from person-to-person. Lacan would develop this further by suggesting that signifiers are dependent upon other signifiers for meaning: forming a chain of signification to which meaning constantly relies. Derrida would destabilize language further by stating that written language is always deferring its meaning, suggesting that one can never settle on singular interpretations. These are all well-known ideas to most students and scholars studying language. It would therefore seem that any endeavor to wrest meaning from language—even in the domain of science—must confront these same issues.

Since those that study language suggest that a discrete, one-to-one approach to interpretation language is not possible, implementing computational processes that are founded on discrete principles would seem problematic. I argue that it's not useful to coerce language into a fixed quantifiable value when the very nature of language repels it. Sentiment analysis in natural language processing requires a new approach that recognizes the slippery nature of language. Using stock market sentiment as a case study, the new approach I want to now propose borrows in part from the study of affect theory.

Many instances of affect theory, particularly in the works of Silvan Tomkins, seems to reproduce the same discrete categorizations that I argue are at odds with language itself. However, there is one aspect of affective response that I believe may be used to help unravel that subjective metric in stock prices: the concept of “texture.”

The notion of texture as it relates to affect theory derives principally from the work of Eve Sedgwick in her work *Touching Feeling: Affect, Pedagogy, Performativity* and Renu Bora’s work on texture study (specifically his essay “Outing Texture”). Introducing this concept, Eve Sedgwick writes:

> [T]o perceive texture is always, immediately, and de facto to be immersed in a field of active narrative hypothesizing, testing, and re-understand of how physical properties act and are acted upon over time. To perceive texture is never only to ask or know What is it like? nor even just How does it impinge on me? Textural perception always explores to other questions as well: How did it get that way? and What could I do with it? (13).

Texture can be seen as a combination of both materiality and active subjective engagement. Since stocks are a complex intangible property—yet tied to actual monetary value—the notion of "materiality" requires slight revision. I would argue that the material properties of a stock are the fundamental empirical values that may be attributed to a company. As previously mentioned, this includes monetary values such as assets, income, debts, property, and all properties that can be reduced to an exact value. We understand that a stock price is based at least partly on these fundamentals, and the other on a more subjective metric. Interestingly, the notion of texture also has a similar relationship to both materiality and subjectivity. Texture, including its absence in the form of smoothness, what Bora calls "a type of texture and texture's other," is a subjective property of materiality (Bora 99). One cannot calculate texture in the same way that a chemist might be able to breakdown and measure the composite elements that make up an object. Yet, texture is still an important property relating to how humans interact with material objects.

Since textures are subjective they are not universally agreed upon. One might suggest the texture of corduroy is rough, while another might claim it's soft and smooth. Additionally, a texture like "rough" refers to large spectrum of "roughness" and its perception relies exclusively on the individual's sensitivity. Sedgwick also makes an important observation between the affective connections of "touch" and texture. The sense of touch, a means to measure texture, is an emotive gesture, insisting that there is "a particular intimacy [that] subsists between textures and emotions" (17). Even in the very word, "touching," she shows there exists an embedded double meaning pointing to both emotive response and tactile sense; the same can be found in the word "feeling" (17).

In relation to stocks, I believe the same principles inherent in texture observed by affect theory may be applied to analyze subjective influences in the stock market. Rather than try and determine "sentiment" in a stock, and reduce it to fixed binaristic characteristics, I believe a better approach would be to try and determine a stock's "texture" at a given time. In theory, the idea is to snap-shot a stock at a given moment, aggregate all language data available at that moment and compress, congeal, and collapse this data into some pattern. This texture would not be definable in itself other than being simply a “texture.” The challenge will be to find a way to capture an imprint of this texture, store it and catalogue it. Once sizable catalogs of different textures are available, comparisons can be made and incidences observed where this textural imprint reappears throughout the market across different stocks and across time.

Textural imprints itself are not enough to make decisions on the stock market. There must be both historical data of stocks available as well as historical language data sets. Databases are already available that provide stock data going back several decades, however, Twitter and Facebook have not been around long enough to provide enough historical language data sets. While these social networking services continue to amass more data, I believe one may be able to use historical news headlines as a language data source to draw correlations. Yet, the more data sources available the better the resolution and clarity of the data.

The vast amount of stocks eligible for trading on the market, and the millions of rows of longitudinal data that can be incorporated into analysis opens possibilities for modeling and testing various hypotheses— making corrective adjustments as needed. This very principle is precisely the process we use to observe texture in the natural world, a process of being "immersed in a field of active narrative hypothesizing, testing, and understanding how physical properties act and are acted upon over time" (Sedgwick 13).

In conclusion, this theoretical approach to measure subjective impact on the stock market works alongside the premise that language cannot be interpreted directly. Instead of forcing words and semantic elements into fixed categories, this approach allows language to create its own container; its use will be in finding instances of recurrence and using historical data to make inferences of how different textual imprints have different impacts on stock value. The next step in this process will be in implementing and testing this hypothesis: specifically writing code to work through the challenges that have been identified and either confirm or deny whether texture can be used as valid metric to gauge sentiment in the stock market.

Works Cited

Bora, Renu. "Outing Texture." *Novel Gazing*. London: Duke UP, 1997. 94-127. Print.

Chakoumakos, Rowan. *Automated Market Sentiment Analysis of Twitter for Options Trading*. Staford University, n.d. Web. <http://cs229.stanford.edu/proj2011/TrusheimChakoumakosYendluri-Automated_Market_Sentiment_analysis_of_Twitter_for_Options_Trading.pdf>.

Kern, Eliza. "Facebook Is Collecting Your Data — 500 Terabyes a Day." *Gigaom*. N.p., 22 Aug. 2012. Web. <http://gigaom.com/data/facebook-is-collecting-your-data-500-terabytes-a-day/>.

Olanoff, Drew. "Facebook Announces Monthly Active Users Were At 1.01 Billion As Of September 30th, An Increase Of 26% Year-Over-Year." *TechCrunch*. N.p., 23 Oct. 2012. Web. 23 Oct. 2012. <http://techcrunch.com/2012/10/23/facebook-announces-monthly-active-users-were-at-1-01-billion-as-of-september-30th/>.

Sedgwick, Eve Kosofsky. *Touching Feeling: Affect, Pedagogy, Performativity*. Durham: Duke UP, 2003. Print.