

**BUSINESS INTELLIGENCE TO AID DECISION MAKING FOR INSTITUTIONAL INVESTORS IN THE CYBER SECURITY SECTOR**

**Supervised by:**

**Dr. Timothy Summers**

**Submitted by:**

**Sanjna Srivatsa**

# Abstract

The world of finance and investment is moving toward highly data driven methodologies aiding in decision making. But investors complain that these methodologies are not agile and don’t tend to keep up with current trends of the market. Business Intelligence has been helping corporations with decision making since decades. Our research cascades the use of Business Intelligence in the world of investment and decision making. The idea is to build a comprehensive, informative and easily digestible indicator for institutional investors that is agile and self-optimizing. This paper demonstrates the prototypical effects of this model in the field of cybersecurity. We have incorporated a computational, optimization and a corrective module in the operational model. With a multi model database structure storing data from various input resources, and Apache Kafka queueing data we have used improved algorithms and workflow design to derive a robust and intelligent system. JavaScript’s D3 is used for powerful and efficient visualizations. This model can be easily integrated with other reputation index computation systems, as it provides normalized multiple standard representations.

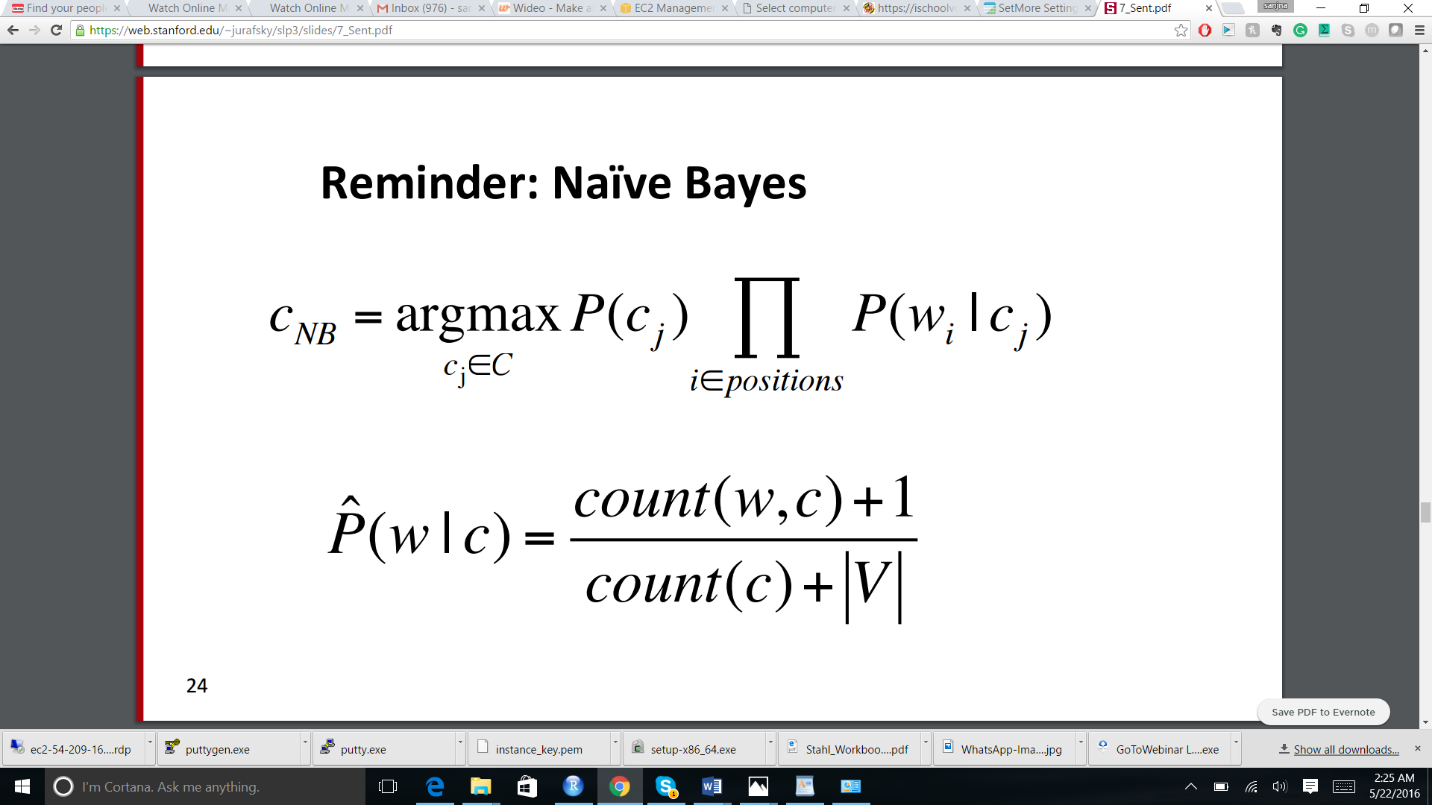
# Entity extraction

In news and technical articles, as in general user reviews we do not have just one object that all comments are directed toward. We have multiple identities and pivoted statements. After performing repeated testing with most popular algorithms we identified that this is not being addressed robustly. Therefore our model tackles the issue of correctly identifying all the entities in the news articles to let us correctly identify and link the sentiment value with the entity.

# Sentiment Analysis

Sentiment Analysis is an important part of our model, and it is vital to have accurate analysis in order to maintain the integrity of our indicator. During the course of our research we encountered various sentiment analysis algorithms which are used widely these days. After rounds of testing we found glitches in these algorithms which we wanted to tackle with our algorithm. There were problems with sensing sarcasm, detecting double negatives and linking adjectives with the appropriate subject. One specific technical oversight was the lack of purposeful modelling or specificity in our vocabulary for sentiment analysis. The word small may mean to be good and compact in terms of procedures and manual processes but might mean completely differently in terms of hotel reviews. We will induce a sense of specificity particularly in terms of cyber security (this will help with identifying the overall tone of extremely technical news resources carrying rich information).

We propose using a system of unigrams, bigrams and trigrams to aid creation of our corpus. To correct for this irregularity which occurs in cases of double negatives we will deviate from the traditional approach. We will scan the sentence for all the words (there is debate about if to use all words or only sentimentally strong words, but in our case all works work better because every line of the news has something new to reveal).

Moving on let’s talk about the corpus for the sentiment analysis. We will use a collection of historic news articles and cyber specific news resources to train and initially test our model. Once we have built the corpus and the algorithm understands polarity in sentiment specific to cyber security, we can test it regressively on real time data. Each time a word from the corpus is encountered a counter is activated and counted up a number 1 through 5 depending on the intensity of the polarity of the word. This is called the frequency counter. We have added a feature which scans 3 words before and after each word encountered in the corpus of a certain polarity just to verify the entity it is associated with and also to check the context it is being mentioned in. The structure of the algorithm would scan the sentence from right to left. This is because there is a higher chance to encounter the subject before the object of the sentence and hence more accurately identify the object of the sentiment. With our entity extraction our algorithm is also equipped to handle the occurrences of multiple entities causing pivots in sentiments in sentences. Also, the close occurrences of words from the corpus like “extremely high levels of vulnerability to smtp attacks” are tackled. When we encounter so many words with such varied polarities all in one sentence, it is important to identify the context and the entity it is driven toward. The word high doesn’t mean anything until we scan the word ‘extremely’ before it to know the polarity of the word high and the word vulnerability to know the adjective of the ‘the smtp attacks’ being addressed. We will use Boolean Multinomial Bayes classifier among others (like SVM) as a classifier to help figure out what word from our news segment fits into what bag for further sentiment analysis.

For example, “Earlier today, the CEO of Cisco mentioned at a press conference that the competitors are going to hate to love how secure their clients are going to be with the launch of their new product.” Our algorithm can identify that this is a sentence aimed toward Cisco’s new product and there strong positive spin to it.

# Why cybersecurity?

A cyber breach can break the toughest of companies, so defense against cyber-attacks can make or break a company. All companies are investing heavily in protecting their assets from being breached. The cyber security sector is currently in peak boom, so this is the best sector to train, test and demonstrate an agile model. Since all companies are getting into business deals with cyber security companies this is a great investment opportunity to buy into at this time. Our model provides a modular piece that fits in an equation in calculating the best investment opportunity in this sector.

# What is our contribution to the field of business intelligence?

Business intelligence is an up and coming field in today’s data management world. Traditional business intelligence practices talk about looking at business intelligence from an insider’s perspective. Companies hire business intelligence experts to manage company data and provide performance metrics. This model talks about an unconventional approach of looking at business intelligence from a third party perspective. Our data sources are open, public and non-traditional, but effective. The idea behind using this approach is to utilize publicly available data to its fullest. We do not always what transpires within a company but news resources are a pretty reliable resource to try and analyze what a company’s future might look like. That is our point of focus.

# What is the ultimate aim of the model?

We are trying to ultimately achieve building a model which will yield a wholesome and easily digestible indicator of reputation or perception about a company/field from single or multiple complex data resources. This essentially is a data driver decision aiding model which implements the principles of business intelligence. With the implementation of this model for a particular problem statement we will introduce agility and clarity in the decision making process for any level at the corporate pyramid. To demonstrate results of the model we will delve into the cybersecurity space in financial decision making. This model is not an attempt to replace or mimic the currently available reputation indices or stock market indicators, but a comprehensive indictor for the cyber security space. Although there are a number of stock market indicators currently used, when asked institutional investors have reportedly faced trouble finding a way to factor in cyber security during investments. This model is using publicly available data on cyber security companies to compute, optimize, correct and validate an indicator of the current state of companies of the cyber security world and depict what company has a trending future which it later on validates and self corrects.

# Data resources

We have chosen to consider news from Pew review’s top 50 news resources for both objective and world news. News resources are considered to be more or less strictly objective and unbiased, but more often than not it’s not the case. We have a filtering mechanism to segregate these 50 news resources into strictly objective news, world news, opinionated news and irrelevant. We feel world news varies slightly as compared to news specific to the USA, this in turn might affect stock prices or investment decisions. However this seems to be a matter of discretion, so we will maintain a toggle option for an active world news inclusion. As time goes, we train our model to segregate these resources into the relevant labels. We perform opinion mining of the comments section which we do not find in traditional paper news broadcasting. This way we can compare our in the over-all general user opinion and the opinion that the news represents, and any anomalies can be factored for. Next we have Cyber specific news, the top 10 chosen by CIO.com. When we speak about a particular sector, especially one as technical as cyber security, news can vary in content majorly from general news providers and cyber security news channels. Therefore we will use cyber security news to accompany general and world news.

For our optimization model, our major input is data collected from Norse. Norse provides real-time data about cyber-attacks, its type and location of the attack. We will collect data about only the type of attack and collect overall daily statistics about most vulnerable/attacked locations. This data is analyzed and mapped to data analyzed from our computational module. Our computational model covers resources which talk about what each cyber security company is doing, their new products, their innovations, their partnerships etc. We map the type of attacks and location data to which company is actually solving a real-time security issue. We can calculate correlations and mapping subset overlaps to find which companies are best options to invest in, with respect to solving real-time difficulties.

For our corrective model, we use google trends, a simple but effective tool. This data resource talks about what the world feels is important and we can check and correct our focus. Often times the world might be interested in things that we might find not that important. This part of the module collects data for the cyber security space and identifies trends, statistics about topics etc.

For our final validation module, we use stock market data, collected both historically and real time. This will help during our visualization phase when we map both performance of a company (in this case normalized stock price) and our indicator. If we find volatilities in the market, the indicator will be rendered unreliable during that phase. During normal market functioning, our model will check for anomalies in our predicted indicator trends and actual stock price trend. According to the results we will moderate weights allocated for each module – computational, corrective, optimization- in the model.