# Alignment of structured and unstructured data for decision support

Author: Sören Bey
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
s.bey@student.utwente.nl

**ABSTRACT:**

**Since the rising concept of social media, companies try to make use of the information that is provided by these networks. The difficulty of such data lies within their unstructured nature. In order to overcome these boundaries and discover the structure behind such data, managers implement a structuring process that involves other structured data from a company itself, e.g. sales numbers. Managers have bounded rationality, which affects their decision model and ultimately the structuring process. The paper proposes a method that includes an approach to cover the structuring process and enhance the level of relevant structured information. Aside taking into account internal factors and statistical relevance, sentiment analysis plays an important part. The alignment of both data types involves the decrease of redundancies and errors among the process. Additionally, two online tools will in combination with a case prove the feasibility of the method. The case includes information from social media and sentiment estimates on investment moods.**

**Supervisors:**
**Dr. A. B. J. M. Wijnhoven**
**Dr. M. E. Iacob**

**Keywords** Structured data, unstructured data, sentiment analysis, information semantics, decision making

# 1. Introduction

Ever since companies try to improve either a product, a service or a process step through observation. Former founders of ideas behind such measurements were the Gilberts with their motion study (1919) or Taylor, the founder of scientific management, in 1916. Nowadays companies do not only observe their employees in order to improve the services and products, but they also take into account external perspectives. These external views can be surveys or verbal feedback meetings for example. Due to the vast emerging technology, customers could be asked more efficient for their opinion. Platforms such as Facebook or Twitter provide the place to collect these responses. So far firms already keep track of their sales and have initial ideas on how well a certain service performs. Social media provides the opportunity to create a more complete set of data for more relevant information through larger samples.

In the early stages of observation, the process to obtain data was rather difficult. Nowadays technology provides tremendous amounts of data. These captured data is called "Big Data" (George, Haas and Pentland; 2014). George et al. (2014) define big data as data acquired from multiple sources such as internet, user-generated or mobile interaction. The huge amounts of big data are difficult to analyze due to their complexity. Through common understanding it is already known that firms have data available, which is supposedly already organized and specified to their decision needs. These two different types of data, structured and unstructured, can be combined to create a more relevant and complete set of information. Within the complexity of information, managers consider several criteria for higher quality decisions (Kangas, Leskinen and Kangas; 2007). This leads to the central research question of this paper: "By what method can structured data be aligned with unstructured data for an improved service-decision-support?"

The goal of this paper is to suggest a method that consist of several elements and allow the feasible alignment of the found data. While the academic literature offers solutions to structure data, it fails to deliver an application within the area of social media in the context of internal and external alignment. As for firms, the method has the means to enable the production of more relevant information. A major aspect is the interaction with the customer without engaging them, simply by analyzing their social media data. These results can be summarized to target specific customer groups for a more specified service. Several sub-questions arise through the practical implementation, for instance "how can unstructured data be structured" or "how are the different types of data aligned, so a firm can make use of additional information". The following section will try to answer the questions and conclude with a real-life example to prove its feasibility.

The idea of the method is plain simple. The structured data and unstructured data can be aligned for more relevant information that achieves higher quality in decision making. The unstructured data is provided in form of text and extracted from social media. On the other hand the structured data is made available by a firm and consists of variables that are of importance for their decision-making. So the method of alignment from structured and unstructured data can be seen as a multiple-point plan. Hereby each aspect has its' own difficulties in order to produce more relevant information. First, the structured data can be considered the starting point for the process and provides the initial idea in which direction the method is applied. Once the structured data is gathered, the decision maker has the variables necessary to make a decision at his display without considering external data, yet. From this point on forward the firms' objectives are known and the manager can start to screen social media for his pre-defined variables. Menon and Pfeffer (2003) discuss the importance of external data that is needed to complete the existing internal data. In this case the social media data serves as an addition. However, the process also works the other way around in order to define areas that the manager has not yet considered. This means that external information such as data from social media provides ideas to a firm where customers might suggest change and improvement. Hopkins and King (2010) call these the supervised and unsupervised approach. Further, these data are relevant due to the fact that the firm has not been aware of the problem areas. These newly aggregated information enable the sharpening or change of the decision. Several theories will be elaborated in part five – unstructured data - to show how relevant issues in unstructured data are discovered. Once these steps are achieved the data can be aligned to create a new data set. In the last step, knowledge can be derived from the newly formed information.

The aspects together add up to a new method as a whole. The section that defines the idea of structured and unstructured data are part of the method and can be considered sub-processes. These again are relevant to follow up with the alignment of both data-types. While, the structured data as summarized in the fourth part of the paper, provides an initial starting point, the unstructured data needs to undergo a process in order to add a structure to it. It is advisable to consider the different types of decisions a manager needs make when applying this method since the method is time consuming. Three different general types of decisions are commonly known. These three types are the operational-decisions, tactical-decisions and strategic-decisions.

In the context of the method it is necessary to distinguish between those types and since the level of complexity of the method is high it is rather feasible for tactical- or strategic-decisions. These need high quality in decision-making for long-term projections. As soon as all these steps are applied, the (un-)structured data can be aligned with the structured data.

The data collected over a five day period are of qualitative and quantitative nature. Further in the process these are summarized to create conclusion that can be interpreted by the manager. The structured data is made available from an online website such as Opfine.com and a banks' analyst. As for the unstructured data, the social media platforms such as Twitter or Facebook provide the data sets. These are limited through queries, since not every single post is relevant for the research. The tool offers the possibility to create queries that enable a search for tweets. Each source was monitored for five days and noted in an excel table. The first feature that was observed, was the mood-news-estimate on average per day. Additionally in social media, the sentiment of the postings and most frequent used words were noted. Together the data provide a completer set of information.

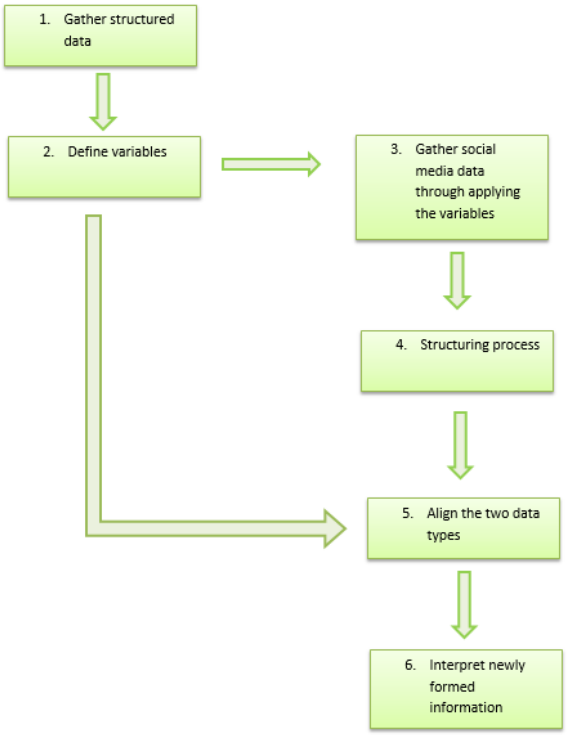



Figure 1: multiple-step-method

## 2. Structured and unstructured data

In the context of the method the paper first focuses on the structured data. The name itself already offers a small hint on the type of data that is considered here. However, a clear distinction from other data types, such as semi-structured or unstructured data, needs to be made. Sukanya and Biruntha (2012) state that structured data is "(…) useful information (…) such as classification, clustering, visualization and information extraction". Marvasti et al. (2013) comply with the previous definition of structured data by Sukanya et al (2012). This indicates that patterns are recognized and hereby enable the solution of more relevant information. Furthermore, Baars and Kemper (2008) understand data as "(…) assigned to dedicated fields (…)", which means that structure is a form of data that can directly be used to extract knowledge (Bellinger, Castro and Mills; 2004), without taking any action towards structuring the data. This includes search queries and the aggregation of the data to entities. It is important to mention that structured data enhances efficiency of search and knowledge extraction.

Structured data can have several forms. These can be understood as dimensions where different structures are applicable. These depend on the circumstances that surround the cause of the data's' existence. In decision support the dimensions are closely related to the type of decision that needs to be made. Each decision requires individual dimensions. Another dimension is the implied versions of a database. Here a database has an existing structure, which is applied to the incoming data. Through understanding this structure it can be concluded with the third dimension. In return the pragmatic approach applies its own structure to the database. Instead of being divided into the different entities of the database, the base is set up according to the data that is provided. This can be understood as a bottom-up structure.

Structure can appear in different formats either defined through the decision model, on the database it is secured in, or the data has its´ own structure. In the first case managers and decision-makers (Azapagic and Perdan, 2005) define the variables and characteristics that are of relevance for the choice that they want to make. When the managers point out the characters they need, they most likely make use of aspects that are of high significance for their decision. Since they described the important aspects of their decision, the structured data is defined to their objectives. However, it is necessary to address the importance of this point in the method since it becomes relevant again in the structuring process of unstructured data. The variables provide the possibility to pre-structure the unstructured data. Hereby, data mining uses these characteristics to initiate the process of structuring data.

According to Sukanya and Biruntha (2012)'s definition of structured data, the unstructured opposite is not structured but rather chaotic. Each data set is considered individually and has not been assigned to any fields (Marvasti et al.; 2013). No patterns or clusters have been detected and no relationships within the data have been established. Rusu et al. (2013) considers webpages or general information available online as unstructured data. Not all information or features available on a website are useful for the decision maker. As for the dimensions of these data sets, they exist but each data set has its' individual dimension and no consensus is reached. When addressing the format, the data can be numbers or words placed aside without being related. These data now also have the same origins. First the unstructured data that comes from a database. They are unstructured in itself and through the decisions that they are required for. In order to derive more relevant information from it and create general proportions that enhance decision making in services, structure has to be added. Structure can be discovered by applying different methods that enable the addition of relevance to different aspects of a data set. This also includes counting frequencies or take into account the sentiments. Hereby it is difficult to take into account the emotions behind what has been said. For example a post on Facebook, it is hard to define whether the author intents to display a certain degree of sarcasm or not at all. Of course this underlies certain limitations such as the author is the owner of the Facebook-page and not a third individual. In the following some methods that add structure to unstructured data are considered and explained.

## 3. Structuring unstructured data

### 3.1 Unsupervised method

As for the structuring of the social media, Hopkins and King (2010) distinguish between supervised and unsupervised approach. The supervised method includes the pre-defined variables, as mentioned earlier and applies these to the "new" data in order to define entities and relations. The unsupervised method includes a structure that arises from within the data without any pre-definition of relations or entities. Hereby frequencies and statistical relevance are essential. These frequencies can only be counted after the data has been structured since all sets need a common nominator. Madhusudhan and Rao (2013) developed an eight-step plan to specifically structure text data. The process starts with gathering the necessary elements, eliminating irrelevant information – data redundancy - and proceeding it through a warehouse. This is considered the

preparation phase of the method. The concept of data redundancy becomes important again in the actual alignment of data (Kolahi and Libkin, 2010). Once these steps are accomplished the authors suggest the definition of features that allow the mining and pattern detection. Following upon the action the interpretation and elaboration takes place. When the authors refer to features they imply the definition of variables. The variables are decision-criteria, which the data sets are based on in order to accomplish valid results for the managers. From these results a new database can be created. The discovery of relationships among the sets is possible. This step requires the creation of clusters and summarizes the data sets in their respective category. Finally the authors conclude with the conversion of the newly formed data set into an interface. The interface is important to display the information and make changes visible for the managers.

The so proposed eight-step-plan enhances the use of unstructured data in an efficient and simple way. However it lacks the depth of certain aspects, such as the definition of features and the possibility to take into account the variables of the decision model. Therefore it is necessary to consider other methods that enable the development of their idea to a more complete method. Rusu et al. (2014) follows a similar approach. The authors suggest the "KDD" process. The knowledge discovery in database process contains five steps to extract information. From selecting the data, over pre-adjusting and transforming to data mining in order to result in elaboration and interpretation. The authors refer to the semantics when creating a structure. They create rules to classify the sets and make use of an algorithm to achieve the desired outcome. In the next paragraph the paper will consider a supervised approach and follow up with the detection of statistical relevance.

As said above the suggested procedure misses a sophisticated explanation on the definition of "features". Never-the-less, their ideas do serve as the outer framework in the structuring process. The part now considers the explicit process of detecting relevant aspects within the data.

Baars and Kemper (2008) imply and interchange between the extraction of knowledge from structured and unstructured data. Aside their first and third layer, the second "logic layer" contains the systems to apply structuring techniques, i.e. "query based access". While these techniques provide the possibility to derive real-time information without pre-defining variables they do not deliver highly meaningful information if phrased poorly. The idea of the interchange introduced by a manager, i.e. through queries, can extract knowledge but also include new aspects in the system to enhance the structure. In application it shows that through developing variables, by common sense, an initial structural process within the database occurs. These are the so called decision variables. This is important because it accounts for a large type of existing data. However, it does not create clusters for all available data. According to Galbraith (1977) managers are unable to process all available information and therefore cannot define all available variables. This results in huge amounts of data that has no use due to the limited defined queries.

Following from the previous method by Kemper and Baars (2008), we now move towards the unsupervised method in order to structure what not yet accounted for. Marvasti, Poghosyan, Harutyunuan and Grigoryan (2013) mentioned a tool to identify patterns by involving the statistical relevance. Hereby each data set is transformed into an "event", which allows the authors to define it by its characteristics. These for example are place of creation and time. It is important to mention that each data receives their individual amount of characteristics to ensure the account of all relations. Furthermore, if one "event" has a unique characteristic it is considered irrelevant and can be eliminated. This offers the opportunity to generalize and summarize attributes. The authors identify five aspects that are relevant in an analysis, namely "critical node" that shows more often occurring events.

The "root", which displays events that have an impact on others only, and "critical path" that shows a consecutive relation between data with high relevance. In addition to that the "extreme path" displaying superior probabilities and "critical sector" that provides information on assembled connections. How does this help to structure data? In the case of unstructured data from social media, each statement can be put into perspective of one another and determine its' importance in the context of a whole. This will result, according to Marvesti et al. (2013), in a developed graph with the most relevant information. Frequencies become important in this case. The more often a certain attribute is mentioned it increases its' relevance in the entire dataset. The idea behind this procedure is rather complex and it involves a vast amount of calculations that are very time consuming. Never the less it can be more accurate than the following method that includes a sentiment analysis, but that remains to be proven. Brun (2011) conducted a research that aimed at extracting opinions on a product from a website. Hereby it was necessary to put attention on the subject of semantics and controversies in the statements. A relation between this work and the one by Baars and Kemper (2008) can be made. While they focused on the influence of pre-defined queries and initiations of the structuring process by managers, Brun (2011) considers a self-developed "lexicon". The initial ideas, created through queries, are related to the limits of the manager and their decision model. This lexicon contains polarized phrases and verbal expressions that are applied in the structuring process in order to create a stable structure in the available data. In application this means the definition of, especially social media, words and interpretation of facial replicates in form of a smiley (☺ is not ☹). Through defining these, the researchers were able to detect and summarize relationships in the data. This concluded in the display of more relevant information.

In order to enhance the quality of such an analysis the above mentioned lexicon needs to be thoughtfully created. This can be done through interviews of the decision-makers and considering synonyms. The author identified six different connections within the semantic analysis that are important. These are stated characteristics in a posting such as "large" or "small", personal expressions such as "I" or "we" and aspects responsible for judging the service. This can be "good" or "not happy" for example. Furthermore, the analysis allows the applicant to put the expressions into context of the entire posting. In this case the analysis creates a definite relation between the personal pronoun and the actual product or service. This means, connect the individual to his opinion on a certain issue. Again this method also has its' limits. The possibility of generating biased information is strong through mal-interpretation of certain phrases and expressions by the manager.

Another aspect that needs to be mentioned briefly, is the redundancy of data. Especially in social media it is issue since not all texts or posting that are made are important for

a distinctive decision. With the goal to overcome these difficulties the method by Morris and Rob (2007) is shortly introduced. The authors engaged the topic of database design within their research. They distinguished between data inconsistency and anomalies. Inconsistency arise while changes take place that are not realized in the data. Anomalies consider entire entities that can be assigned to multiple unique customers for example. A change in customer status does not suggest a change of the entity entirely but only an update for one customer. In order to overcome these difficulties the authors suggest a thorough database management. The so called "DBMS"-concept contains the rules for the integration of data, as mentioned in their case. However, this can be applied in the above mentioned alignment process simply through concurrent control. Concurrency control was introduced by Bernstein, Hadzilacos and Goodman in 1987. The concurrent control is a method that decreases the amount of biased data. An explicit definition of the concept can be acquired in their book, but within this paper, it exceeds the scopes of the research. Through applying a thorough database management, the redundancies can be reduced. The now accomplished cleaner data displays more relevant information through the increase of quality.

## 3.2 Step of alignment

In the last step of the method both types of data can be aligned. Instead of integrating or including the unstructured data into the structured data, they can both be combined into a new form of data structure. Each data set can be imagined as to separate excel sheets that can be combined to one. This enables the aggregation for an appropriate summary of the data, which in return support a higher quality of decision-making. So the alignment-procedure is rather simple. The two data types complete one another. Here it is essential to know that the two data types can alter each other, relate to the variables of the other or lead in combination to the creation of a new entity. Furthermore, once this is accomplished, the data must be checked for double entries, redundancies or missing sets. While, double entries and falsified data are more important and have a larger effect upon the decision, the missing entries are either not yet available, which means they are not really missing, or do not exist. Larger samples decrease the percentage of redundancies. Of course these applications can be used to consider an improvement in the selection process of the data during a previous gathering data step. An improvement in this step at this point would mean a direct increase in quality.

Following from that the information can now be displayed in an interface. The interface is meant to simply display proportions and provide access for the manager to interpret the findings. This can either be an excel sheet or table for instance.

## 3.3 Appropriateness of the method

So far the method has only been explained theoretically. In order to prove its feasibility the method will be compared to already existing online tools. Two tools are considered. Each one accounts for various aspects of the method. These two chosen tools mirror the structuring process and access to the structure of data. First, the free online tool, see figure 2 below, that focuses mainly on bare numbers and statistics across several social media platforms. For the analysis the website includes Facebook, Twitter and Youtube for example. Aside these, the tool intends to determine the influence the posts have, through taking into account the number of people that have seen or commented upon these. Further a simple analysis is made of the nature of the postings. These can either be positive, negative or neutral. The major aspect of this tool can be seen within the variety of numbers it displays. The output of the example search term "bachelorthesis" was not only the postings made per day or month, but also on which platform it was made. As mentioned above it also indicates the level of influence a post or video has. In order to account for the measurement of the difference between low influence and high influence, the number of views is relevant. Hereby a video with only a hundred views was considered to be of low influence while others, exceeding the thousand views were considered highly influential. As for tweets, the more followers a tweet has the more influence it has.

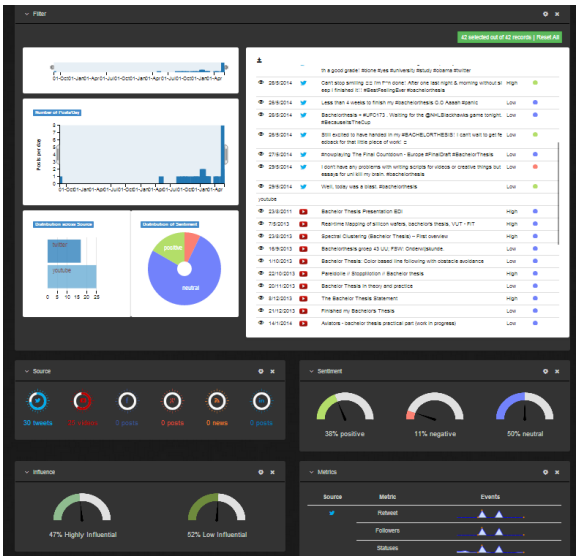



Figure 2: screenshot retrieved from www.sentiks.com

With the focus on these statistics the tool mirrors the above mentioned structured data and analyzes the statements of the users according to their sentiment. However, it is not entirely visible how the tool distinguishes between the different natures in the sentiments. Therefore, the tool only partially accounts for the structuring process. The legitimate choices of the variables that decide on a certain structure are not visible.

Another tool that is offered for free provides a more sophisticated approach to the analysis of social media. It especially considers the statistical importance of the relations between the different words that are used within the statement. The online tool searches for a term, chosen by the user and determines their origin. Once this is done the tool's algorithm identifies the nature of the posting and clusters them as visible in figure number 3. In addition to the basic difference of positive and negative, the tool tries to identify the explicit emotion behind a tweet.

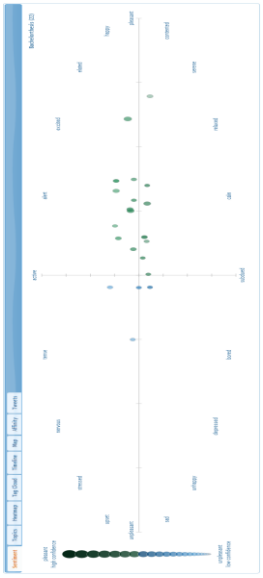


Figure 3: screenshot retrieved from http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

In comparison to the previous tool they offer the so called "Tag cloud". The feature displays the main words used in the postings and distinguishes between good or bad characteristics. These words are part of the ANEW dictionary, which is meant to create a universal standard in the study of emotions. Figure 4 below shows the words related to the search term "bachelorthesis". This finding can be related to the topic of predefined variables above, including the terms defined by a manager and independent lexicon.

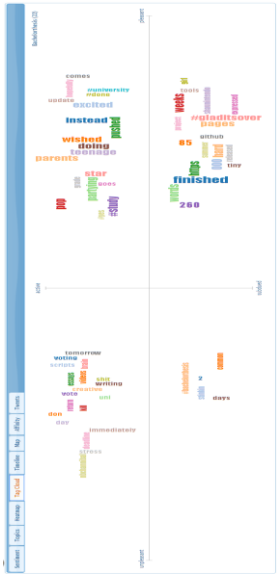


Figure 4: screenshot retrieved from http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

The tool already takes into account the approaches of supervised and unsupervised methods. However, when having a closer look at the different states of mind the tool considers and tries identify these. It also still has difficulties as mentioned earlier, to understand emotions. As for many systems it is difficult to distinguish between sarcasm and honest opinions. In this case it fails to deliver high accuracy, but never the less enhance the quality of the information. Aside determining these words and clustering these according to the statements nature, the tool takes another step and involves the statistical importance. Below two different tweets are displayed directly from the tool itself (Figure 5). The top post is considered to be a positive and the lower post is considered to be a negative statement. The bold formatted words at the bottom of each, are the words relevant for the webpage. These terms are the most meaning full within these tweets and determine their nature.

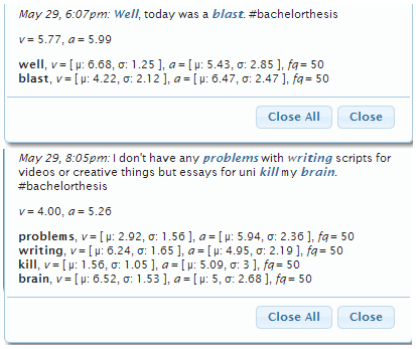



Figure 5: screenshot retrieved from http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

In order to analyze their nature, as said earlier, the creators make use of the ANEW dictionary – Affective Norms for English Words (Bradley, Lang and Cuthbert, 1997). The ANEW dictionary offers numbers and ratings on English words according to their dominance and pleasure or arousal, with the goal to create a universal standard within the study of emotions. In the context of this tool, statements with less than two words that are available in the ANEW dictionary are not considered and eliminated. In the example postings above, "well" is more important than "blast" and therefore received a higher mean. However, while receiving the higher overall mean according to the dictionary, the level of arousal is lower than the "blast"-value. These values together determine their place within the grid. The higher the "v"-value the more likely the statement is to be pleasant and the higher the "a"-value the more active the tweet is.

Through involving the ANEW dictionary the tool involves the aspects described in the structuring process. It can be considered an independent dictionary that not only accounts for the postings but also involves the statistical relevance of the different words to form an overall value. In return it can be said that the proposed method to structure unstructured data, is applied within this web-tool. Furthermore it is necessary to mention that the ANEW dictionary offers a different kind of importance measure than previously indicated. It is similar to the method mentioned by Marvasti et al (2013). Not only do they agree upon the fact that events that have one unique characteristic are excluded, but also the more often it is mentioned it becomes more relevant. So to say, while the ANEW dictionary is based on a huge survey where participants needed to associate the characteristics with an event for a general applicable dictionary, Marvesti et al. (2013) create a new dictionary unique to each subject. Both ideas aim at an equal goal, namely to standardize the meaning of words to make them measurable. This means that the method is appropriate in terms of the process to structure, which is a sub-part of the entire method, but still is no indication of the overall feasibility. In order to show

its' feasibility, a practical case on decision to invest was created. The elaboration on this topic follows in the next paragraph.

# 4. Case study: investment decision and sentiment

## 4.1 Methodology and data collection

This section studies the question: how likely is an investment in the near future? In order to derive at a conclusion, the decision model needs to be understood. Bellman and Zadeh (1970) already referred to difficult "fuzzy" boundary definitions and call for sharply defined objectives. Due to the simplicity of the case, the decision model is based on four different factors. These are return on investment per share estimated for the year 2014, change in performance over a 12 months period, overall average and individual mood-level, and Twitter sentiment. While the first two factors are provided by a bank, the average mood-level is an estimate of an online website for long-term projections. The website opfine.com, displays several different online newspaper articles that influence the attitude to invest. It is distinguished between positive and negative articles. Positive articles increase the mood to invest, whilst negative articles decrease their calculation of the mood-level. The result is an estimate of the mood-level every hour, which show radical changes during the day. So in order to be relevant for the case study, the average per day is considered.

As previously indicated, the structuring process is rather time consuming. The online tool for the social media sentiment analysis covers most of the aspects that are involved in the procedure and therefore serves as a feasible assistant. In order to receive appropriate information, each company name can be used to create a query and retrieve data from Twitter. Further, these are specified, due to the fact that the social media offers more postings on topics that might be irrelevant and the case only focuses on investment decisions. This was done by adding three different terms to each query. These were "invest", "buy" and "recommend". They were chosen as feasible to add because it limits the amount of post. For example many people used "#invest" to publish their tweet. In addition to that each term was selected after a short verbal interview with a bank director that mentioned these as most important during day-to-day operations.

The sentiment analysis of the Twitter data that is offered by the online tool covers several dimensions of emotional states. Instead of focusing on each one of them, these are limited to the positive and negative proportions. The results are displayed in the following paragraphs. In total, three different sources are leveraged for data sets and information for case. First, the company portraits, including the ROI and performance measurement, were provided by a professional financial analyst (National Bank AG, 2014). The mood-level estimates are from a professional website, which is limited by the amount of company data and information on its determinants. These were matched with the content from professional analyst and resulted in 15 selected organizations. As for the social media analysis, the tool is somewhat unprofessional, but the methodology of analysis is known. Figure 6 shows a table of the different sources.

| SOURCES | unprofessional | professional |
|---|---|---|
| known methodology | social media data | company portriats |
| unknown methodology | | opfine.com |

Figure 6: data sources

## 4.2 Gathering structured data

In total 15 companies, all part of the Dow Jones 30, were selected to participate in the study due to information that could be matched from the website and the bank's analyst. Figure 7 displays the return of investment per share, estimated for the year 2014 on the current stock price. The variable was chosen because it is universally comparable among all companies. However, they are limited to the respective home currency of their stock listing. The highest ROI has JP Morgan. This implies that it is one of the most attractive opportunities to invest. Chevron and IBM follow with the second and third highest ROI, with respectively 8.64% and 8.96%. The Coca-Cola-company and Procter&Gamble have the lowest return on investment, with a value of 5.08% and 4.99%. Therefore they seem like the least attractive option for investors within this group.

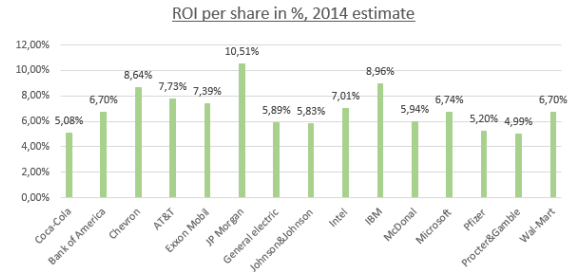



Figure 7: ROI per share estimated for the year 2014

Furthermore, the graph in figure 8 displays the second factor of the decision model: the change of the performance of the stock price for each company within the last 12 months. Evidently Bank of America, Intel and Microsoft show the highest increase, while Exxon Mobile and Coca-Cola show a definite decrease in performance level. Here the Bank of America and Microsoft seem as the most promising opportunities, while the other lack the necessary performance level to be considered an attractive investment. So far these facts are only indicators of attractive options. They are not yet the complete set of information. The results point out that an investment will be likely in IBM, Microsoft or JP Morgan.

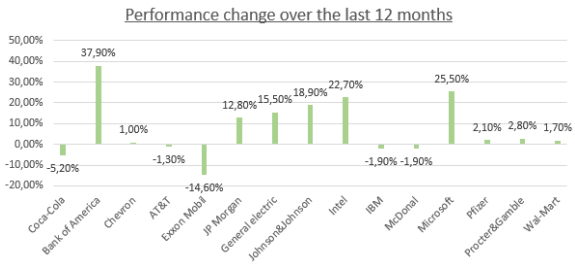



Figure 8: Performance change over the last 12 months

Now, the data provided by opfine.com is considered. When taking a look at the average mood level during the period, some changes can be detected. So far it can be said that the mood-level towards making an investment moves between 55,10% and 62,86% within the five day period (see Figure 9). This means that more than every second person is likely to choose to invest.

The data does indicate special tendencies for investors to take action. It gives an overall idea on the current willingness to invest. Currently it ranges from 55,10% to 62,86%. Aside the news analysis the website does not deliver any other indication or in depth information on the topics. It generalize the data. In this case it needs to be mentioned that the sudden increase between day two and day three is due to the introduction of the penalty interest by the European central bank. Even though the news were negative, the result was a change in all investment plans, instead of saving the money, towards investing into more rewarding possibilities. Overall a positive tendency is visible.

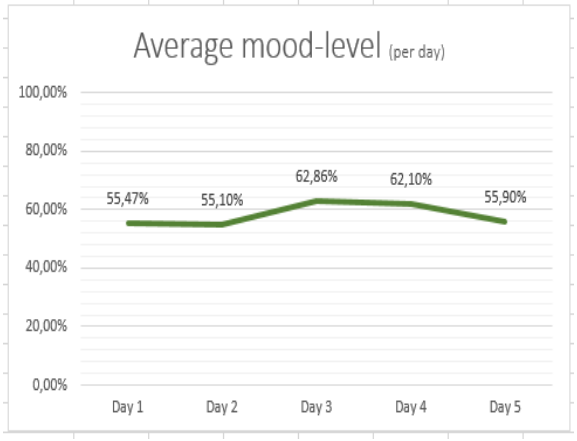



Figure 9: Average mood-level per day over 5-day period

Furthermore, the current-individual-mood-level is displayed in the second column in figure 10. These values are unique and make the ratio-calculation of the overall-average and current-mood level for each company possible. These can later on be compared to the social media positive-negative ratio by executing a two independent sample t-test.

## 4.3 Gathering and structuring unstructured data

Structured data such as the mood level delivers a broad overview on the general attitude of the investors to take action. These data can be specified through including social media. Social media therefore delivers data on the choice of investment. Tweets made on Twitter were analyzed with the previously mentioned tool. As for the company AT&T the results were too broad to gain accurate information and was therefore left out.

The other queries created from the remaining companies delivered results. In total 395 tweets were found to be relevant for Bank of America. From the general outlook it can be said that the spread shows a rather positive association with the company (see Appendix 4). Most of the tweets, above 70%, pleasantly mention the company.

Chevron and Coca-Cola, 356 and 466 tweets in total, show a similar pattern. The results display a more positive attitude towards these companies. In comparison to the Bank of America the tweets are more frequent in the pleasant part of the model, so as the strong relation of the company with positive characteristics. As for JP Morgan the results are centered. Even though many outliners exists, the general sentiment remains centered.

Exxon mobile barely made the cut, since only 49 posts have been found to be relevant. With six negative tweets and 43 positive tweets the sentiment also shows a positive association with the company. Around 12% of all tweets were negative, which is the lowest value among all companies. It might be the most interesting investment option. IBM showed the highest positive-negative ratio with 23.85. Barely any tweets were negative. This means the overall experience of IBM presence is positive. In comparison to General electric or Microsoft, the sentiment ratio of the Bank of America is one of the lowest that have been found. The same outcome has been observed for McDonald. The table below displays all observations (see Figure 10).

| Correrelation | Company | average-overall mood level | Current mood level each company | average-current ratio | Company portrait -ROI % (performance %) | Social media sentiment (positive/negative ratio) |
|---|---|---|---|---|---|---|
| | AT&T | 55,10-62,86% | 70-75% | 1,24 | 7,73 (-1,3) | - |
| | Bank of America | 55,10-62,86% | 47% | 0,80 | 6,70 (37,9) | 2,33 |
| | Coca-Cola | 55,10-62,86% | 80% | 1,37 | 5,08 (5,2) | 4,00 |
| | Chevron | 55,10-62,86% | 130% | 2,22 | 8,64 (1) | 4,00 |
| | Exxon Mobile | 55,10-62,86% | 75-80% | 1,32 | 7,39 (-14,6) | 7,33 |
| | Gernal Electric | 55,10-62,86% | 70% | 1,20 | 5,89(15,5) | 11,50 |
| | Johnson & Johnson | 55,10-62,86% | 81% | 1,38 | 5,83(18,9) | 4,46 |
| | JP Morgan | 55,10-62,86% | 58% | 0,99 | 10,51 (12,8) | 3,00 |
| | Intel | 55,10-62,86% | 80% | 1,37 | 7,01 (22,7) | 4,88 |
| | IBM | 55,10-62,86% | 74% | 1,28 | 8,96(-1,9) | 23,85 |
| | McDonald | 55,10-62,86% | 80% | 1,37 | 5,94(-1,9) | 2,11 |
| | Microsoft | 55,10-62,86% | 79% | 1,35 | 6,74(25,5) | 10,51 |
| | Pfizer | 55,10-62,86% | 71% | 1,21 | 5,2 (2,1) | 8,80 |
| | Procter & Gamble | 55,10-62,86% | 77% | 1,32 | 4,99(2,8) | 9,00 |
| | Wal-Mart | 55,10-62,86% | 95% | 1,62 | 6,7(1,7) | 4,56 |

Figure 10: value – table

Beside the sentiment estimate in Twitter, a relevant part are the earlier mentioned side-topics that deliver additional information on investment opportunities. A list of these can be found in appendix A4. The query of Bank of America resulted in terms such as “resourceful”, “advisor” and “merrilllynch”. This means when considering that as an investment, the bank might not only have the necessary resources, but is also connected to another financial organization. Another example is Chevron. Even though the sentiment was majorly positive, terms such as “pollution”, “energy” and “ecuador” arise and create a more complete picture of the company. An interpretation therefore might be that Chevron focuses on a change in energy sources due to environmental pollutions. Of course these are relevant information for each investor. On the other hand, the query of Coca-Cola and their sponsoring efforts of the world-cup, resulted in rather less relevant sub-topics from an investor perspective.

# 5. Discussion

In the beginning of the paper the initial question was: by what method can structured and unstructured data be aligned? The multiple-point plan suggested in the third section takes a leap in trying to gather a more relevant set of information. In order to prove this the case focused on the determination of whether an investment is likely or not. When reflecting the results, only small similarities between

the different sources, structured and unstructured, can be detected (See figure 10 for results). With an average mood level between 55.10% and 62.86% the willingness to invest is positive. Furthermore the estimated ROI is increased in comparison to previous years. This might be due to inflation but that remains as for this paper, part of future research. In order to compare the results and determine whether a correlation can be found, the ratios of the current-average-mood level and positive-negative sentiment were compared via a t-test with two independent samples (see appendix A6 for calculation). As suggested by the outcome, there is no existing relation between the ratio from the professional data and the ratio by the unprofessional twitter analysis. It is highly unlikely, below 0.005% that the data are dependent. At first this seems as a failure in the choice of sources for data. Since the tweets are captured on one specific moment in time and the newspaper articles are analyzed for long term projections, the interpretation needs to differ between long-term and short-term investments. While the unprofessional tool is applicable for short-term investments, the other is relevant for long-term investors. This means that the sentiment analysis is more helpful for investors looking for investments with a short duration. Other data is feasible for a permanent stock purchase. The goal of this research was to create a more complete set of data and by applying the method, the decision-maker receives information on long- and short-term perspective. In the context this means that the proposed method is feasible with the goal to align the data sets for different decisions considering the same topic.

## 6. Conclusion and further research

The goal of this research was to receive a more complete set of information by aligning structured and unstructured data. As the method has shown in the practical case, instead of focusing only on one source of information adding other sources result in a larger set of information. If the managers follow a highly sophisticated structuring process, the level of completeness of information can be improved. This means that the proposed method is partially already present in online tools and can be used by managers in order to gather more data. As the T-test indicated, there was no direct relation between the professional and unprofessional source of data. The interpretation therefore results in the application of sentiment analysis for short-term and the professional analysis for the long-term decisions.

In the future the side - topics arising from initial queries need to increase their visibility. Counting the frequency of the appearance of certain terms is a more pragmatic approach and does not entirely account for all possible outcomes. In addition to that the influence of re-tweets needs to be examined. While single users simply re-tweet postings from larger firms, they might not share the same level of enthusiasm with the company. A possibility to deliver a solution is the level of influence that each user has. The tool introduced in section 3.3 already shows how this can be applied but lacks a thorough explanation on the determination of the influence level. In addition to that the interpretation of the case needs to be tested. Real investment decisions need to be monitored for their performance based on the information that is provided from different sources.

## Acknowledgements

Special thanks to Dr. Wjjnhoven for the close supervision and assistance during the development of the thesis. Further, credits to Dr. Iacob for participating in the last instance and the grading process.

# References


Azapagic, A., & Perdan, S. (2005). An integrated sustainability decision-support framework - Part 1: Problem structuring. *International Journal of Sustainable Development and World Ecology* , 98-111.

Baars, H., & Kemper, H.-G. (2008). Management Support with Structured and Unstructured Data - An Integrated Business Intelligence Framework. *Information Systems* , 132 - 148.

Bellinger, G., Castro, D., & Mills, A. (2004). *Data, information, knowledge, and wisdeom*. Retrieved from http://www. systems-thinking. org/dikw/dikw. htm

Bellman, R. E., & Zadeh, L. A. (1970). Decision-making in a fuzzy environment. *Management Science Series B-Application Vol.17*, B141-B164.

Berstein, P. A., Hadzilacos, V., & Goodman, N. (1987). Concurrency Control and Recovery in Database Systems. *Addison Weley Publishing Company*.

Bongardt, S., Friebel, M., Lusebring, L., Plakhina, E., Riesselmann, M., Ruegzgar, Z., & Schoendube, C. (2014). *National-Bank Research.* Essen: National Bank AG.

Bradley, M. M., Lang, P. J., & Cuthbert, B. N. (1997). Affective Norms for English Words (ANEW). *NIMH Cen, Stud. Emot. Atten. University Florida*.

Brun, C. (2011). Detecting opinions using deep syntactic analysis. *Xerox Research Centre Europe*.

Galbraith, J. R. (1977). Organization Design: An information PRocessing View. *Organizational Effectiveness Center and SChool*, 21.

Gergoe , G., Haas, M., & Pentland, A. (2014). Big Data and management. *Academy of Management Journal*, 321 - 327.

Gilbreth, F. B., & Gilbreth, L. M. (1917). Applied motion study. *The MacMillan Company, New York*.

Healey, & Ramaswamy. (2014, 05 31). *Visualizing Twitter Sentiment*. Retrieved from http://www.csc.ncsu.edu/faculty/healey/tweet_viz/

Hopkins, D. J., & King, G. (2010, Jan.). A Method of Automated Nonparametric Content Analysis for Social Science . *American Journal of Political Science Vol. 54*, 229 - 247 .

Kangas, A., Kangas , J., & Leskinen, P. (2007). Comparison of fuzzy and statistical approaches in multicriteria decisionmaking. *Forest Science Vol. 53*, 37-44.

Kolahi, S., & Libkin, L. (2010). An information-theoretic analysis of worst-case redundancy in database design. *ACM transactions on database systems*, Vol. 35, article 5.

Madhusudhan, C., & Rao, K. M. (2013 Dec). Proposed Architecture for Automatic Conversion of Unstructured Text Data into Structured Test Data on the Web. *International Journal of Computer Science and Network Security Vol. 13 No. 12*, 110 - 116.

Marianmedla. (2014, 6 20). Retrieved from Opfine: http://opfine.com/

Marvasti, M. A., Poghosyan, A. V., Harutyunyan, A. N., & Grigoryan, N. M. (2013). Pattern Detenction in Unstructured Data - An Experience for a Virtualiuzed IT Infrastructure. *International Symposioum on Integrated Network Management*, 1048 - 1053.

Menon, T., & Pfeffer, J. (April 2003, Vol. 49 Issue 4). Valuing Internal vs. External Knowledge: Explaining the Preference for Outsiders. *Management Science*, 497 - 513.

*Pravisha Technologies Pvt Ltd.* (2014, 05 31). Retrieved from http://www.sentiks.com

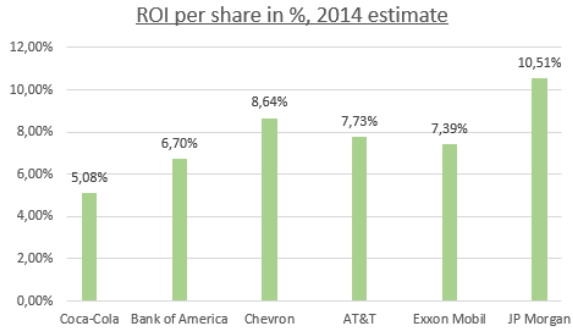
Rob, P., & Coronel, C. (2007). Database Systems: Design, Implementation, and Management, Seventh Edition. *Book*, 13-24.

Suknya , M., & Biruntha, S. (2012). Techniques on Text Mining. *International Conference on Advance Communicaiton Control and Computing Technologies*, 269 - 271.

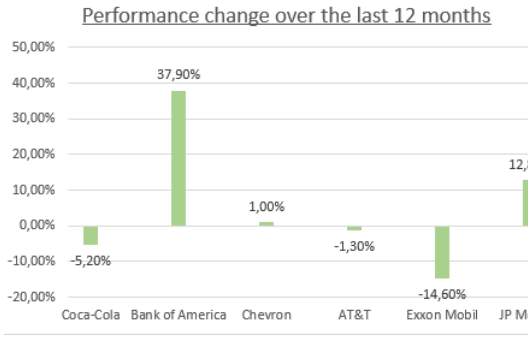
Taylor, F. W. (1916). Scientific management. *Bulletin of the Taylor Society*.

# Appendix:

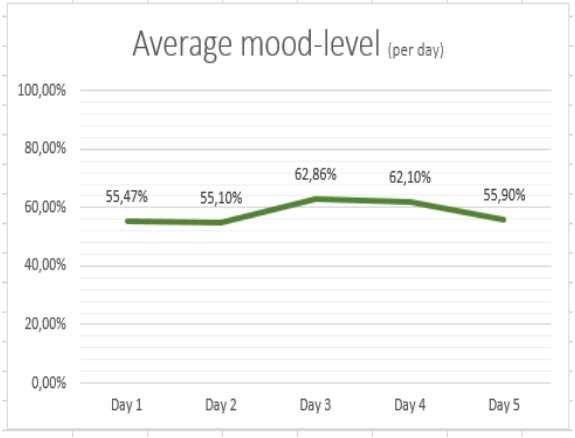
## A1. Graph of ROI estimate for 2014




## A2. Graph of performance change within the last 12 months




## A3. Table of mood-average mood level




## A4. Most frequent side-topics

| Company | most mentioned terms | | | |
|---|---|---|---|---|
| Chevron | ecuador | | Intel | helpshift |
| | energy/changing | | | capital |
| | buy | | | service |
| | pollution/poisoning | | | mobile |
| | | | | customer |
| Exxon mobile | houston | | | |
| | BP | | Johnson & J | flotus |
| | Putin | | | RobinRoberts |
| | Iraq | | | familiessucceed |
| | Oil | | | amp |
| | | | | sudafed |
| Bank of America | resourceful | | | |
| | advisor | | Pfizer | analyst |
| | job | | | astrazeneca |
| | merrilllynch | | | merger |
| | financial | | | deal |
| | | | | lawsuits |
| Jp Morgan | overweight | | | |
| | ranked | | procter &ga | development |
| | complaint | | | manufacturing |
| | top | | | veteranjob |
| | reiterated/quarter | | | technician |
| | | | | job |
| Coca-Cola | open happiness | | | |
| | spain | | IBM | https |
| | Chile | | | business |
| | worldcup | | | wimbledon |
| | share/check | | | cloud |
| | | | | development |
| AT&T | not found | | | |
| | not found | | Microsoft | windows |
| | not found | | | ebay |
| | not found | | | surface |
| | not found | | | owners |
| | | | | pro |
| Wal Mart | times | | | |
| | columnist | | McDonald | sporthumor |
| | amp | | | past |
| | response | | | footballvines |
| | egan | | | morning |
| | timothy | | | drives |
| | | | | |
| General electric | alstom | | | |
| | energy | | | |
| | wins | | | |
| | france | | | |
| | bid | | | |

## A5. Value – table (probability)

| Correrelation | Company | average-overall mood level | Current mood level each company | average-current ratio | Company portrait -ROI % (performance %) | Social media sentiment (positive/negative ratio) |
|---|---|---|---|---|---|---|
| | AT&T | 55,10-62,86% | 70-75% | 1,24 | 7,73 (-1,3) | - |
| | Bank of America | 55,10-62,86% | 47% | 0,80 | 6,70 (37,9) | 2,33 |
| | Coca-Cola | 55,10-62,86% | 80% | 1,37 | 5,08 (5,2) | 4,00 |
| | Chevron | 55,10-62,86% | 130% | 2,22 | 8,64 (1) | 4,00 |
| | Exxon Mobile | 55,10-62,86% | 75-80% | 1,32 | 7,39 (-14,6) | 7,33 |
| | Gernal Electric | 55,10-62,86% | 70% | 1,20 | 5,89(15,5) | 11,50 |
| | Johnson & Johnson | 55,10-62,86% | 81% | 1,38 | 5,83(18,9) | 4,46 |
| | JP Morgan | 55,10-62,86% | 58% | 0,99 | 10,51 (12,8) | 3,00 |
| | Intel | 55,10-62,86% | 80% | 1,37 | 7,01 (22,7) | 4,88 |
| | IBM | 55,10-62,86% | 74% | 1,28 | 8,96(-1,9) | 23,85 |
| | McDonald | 55,10-62,86% | 80% | 1,37 | 5,94(-1,9) | 2,11 |
| | Microsoft | 55,10-62,86% | 79% | 1,35 | 6,74(25,5) | 10,51 |
| | Pfizer | 55,10-62,86% | 71% | 1,21 | 5,2 (2,1) | 8,80 |
| | Procter & Gamble | 55,10-62,86% | 77% | 1,32 | 4,99(2,8) | 9,00 |
| | Wal-Mart | 55,10-62,86% | 95% | 1,62 | 6,7(1,7) | 4,56 |

## A6. T-test calculation

T – Test for two independent samples

Values:

| | Opfine.com (N2) (individual-mood-level/ average-mood-level - ratio) | Social media (N1) (positive/negative – ratio) |
|---|---|---|
| AT&T | 1,24 | - |
| Bank of America | 0,80 | 2,33 |
| Coca-Cola | 1,37 | 4,00 |
| Chevron | 2,22 | 4,00 |
| Exxon Mobile | 1,32 | 7,33 |
| General Electric | 1,20 | 11,50 |
| Johnson & Johnson | 1,38 | 4,46 |
| JP Morgan | 0,99 | 3,00 |
| Intel | 1,37 | 4,88 |
| IBM | 1,28 | 23,85 |
| McDonald | 1,37 | 2,11 |
| Microsoft | 1,35 | 10,51 |
| Pfizer | 1,21 | 8,80 |
| Procter&Gamble | 1,32 | 9,00 |
| Wal-Mart | 1,62 | 4,56 |

H0 = There is no difference between both ratio sets.

HA = There is a significant difference between both ratio sets.

Df = (15-1) + (14-1) = 27

Alpha level: 0.01; two-sided test; equal variances not assumed

Decision-rule: If the t-value of the test statistic is lower or higher than 2.771 or -2.771, H0 is rejected.

N1=14 N2=15

SD=5.685 Mean= 7.167 SD=0.307 Mean=1.337

Calculate the test statistic:

$$SE = \sqrt{\frac{0.0942}{15} + \frac{32.32}{14}} = 1.521$$

$$t = \frac{(x1-x2)}{SE} = \frac{7.167-1.337}{1.521} = 3.84$$

Result: Reject the null-hypothesis because at t = 3.84 the p-value is <0.005

## B1. Sentiment analysis

(Screenshots retrieved from http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/)

Bank of America

Chevron

Coca-Cola

Exxon

JP Morgan

General electric

IBM

Intel

Johnson & Johnson

McDonald

Microsoft

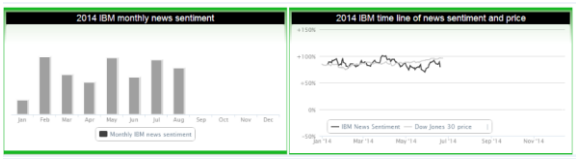
Pfizer

Procter & Gamble

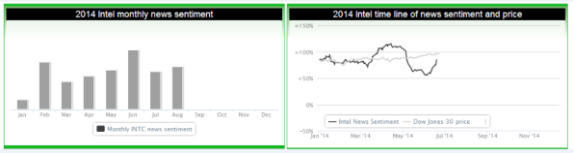
Wal-Mart

## B2. Individual mood-level

(Screenshots retrieved from http://www.opfine.com/)


2014 IBM monthly news sentiment
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Monthly IBM news sentiment


2014 IBM time line of news sentiment and price
+150%
+100%
+50%
0%
-50%
IBM News Sentiment
Dow Jones 30 price
Jan '14 Mar '14 May '14 Jul '14 Sep '14 Nov '14


2014 Intel monthly news sentiment
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Monthly INTC news sentiment


2014 Intel time line of news sentiment and price
+150%
+100%
+50%
0%
-50%
Intel News Sentiment
Dow Jones 30 price
Jan '14 Mar '14 May '14 Jul '14 Sep '14 Nov '14


2014 Johnson & Johnson monthly news sentiment
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Monthly JNJ news sentiment


2014 Johnson & Johnson time line of news sentiment and price
+125%
+100%
+75%
+50%
+25%
0%
-25%
Johnson & Johnson News Sentiment
Dow Jones 30 price
Jan '14 Mar '14 May '14 Jul '14 Sep '14 Nov '14


2014 McDonald monthly news sentiment
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Monthly MCD news sentiment


2014 McDonald time line of news sentiment and price
+125%
+100%
+75%
+50%
+25%
0%
-25%
McDonald News Sentiment
Dow Jones 30 price
Jan '14 Mar '14 May '14 Jul '14 Sep '14 Nov '14


2014 Microsoft monthly news sentiment
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Monthly MSFT news sentiment


2014 Microsoft time line of news sentiment and price
+125%
+100%
+75%
+50%
+25%
0%
-25%
Microsoft News Sentiment
Dow Jones 30 price
Jan '14 Mar '14 May '14 Jul '14 Sep '14 Nov '14


2014 Pfizer monthly news sentiment
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Monthly PFE news sentiment


2014 Pfizer time line of news sentiment and price
+125%
+100%
+75%
+50%
+25%
0%
-25%
Pfizer News Sentiment
Dow Jones 30 price
Jan '14 Mar '14 May '14 Jul '14 Sep '14 Nov '14


2014 Wal Mart monthly news sentiment
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Monthly WMT news sentiment


2014 Wal Mart time line of news sentiment and price
+150%
+100%
+50%
0%
-50%
Wal Mart News Sentiment
Dow Jones 30 price
Jan '14 Mar '14 May '14 Jul '14 Sep '14 Nov '14


2014 Procter & Gamble monthly news sentiment
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Monthly PG news sentiment


2014 Procter & Gamble time line of news sentiment and price
+125%
+100%
+75%
+50%
+25%
0%
-25%
Procter & Gamble News Sentiment
Dow Jones 30 price
Jan '14 Mar '14 May '14 Jul '14 Sep '14 Nov '14