DEGREE PROJECT IN MASTER;S PROGRAMME, APPLIED AND COMPUTATIONAL MATHEMATICS 120 CREDITS, SECOND CYCLE

STOCKHOLM, SWEDEN 2015

Statistical analysis of online linguistic sentiment measures with financial applications

ANTON OSIKA

KTH ROYAL INSTITUTE OF TECHNOLOGY

SCHOOL OF ENGINEERING SCIENCES

Statistical analysis of online linguistic sentiment measures with financial applications

A N T O N O S I K A

Master's Thesis in Mathematical Statistics (30 ECTS credits)
Master Programme in Applied and Computational Mathematics (120 credits)
Royal Institute of Technology year 2015
Supervisor at Gavagai: Jussi Karlgren
Supervisor at KTH: Jimmy Olsson
Examiner: Jimmy Olsson

TRITA-MAT-E 2015:81
ISRN-KTH/MAT/E--15/81-SE

Royal Institute of Technology
SCI School of Engineering Sciences

KTH SCI
SE-100 44 Stockholm, Sweden

URL: www.kth.se/sci

Statistical analysis of online linguistic sentiment measures with financial applications

Abstract

Gavagai is a company that uses different methods to aggregate sentiment towards specific topics from a large stream of real time published documents. Gavagai wants to find a procedure to decide which *way of measuring sentiment* (sentiment measure) towards a topic is most useful in a given context. This work discusses what criterion are desirable for aggregating sentiment and derives and evaluates procedures to select "optimal" sentiment measures.

Three novel models for selecting a set of sentiment measures that describe independent attributes of the aggregated data are evaluated. The models can be summarized as: maximizing variance of the last principal component of the data, maximizing the differential entropy of the data and, in the special case of selecting an additional sentiment measure, maximizing the unexplained variance conditional on the previous sentiment measures.

When exogenous time varying data considering a topic is available, the data can be used to select the sentiment measure that best explain the data. With this goal in mind, the hypothesis that sentiment data can be used to predict financial volatility and political poll data is tested. The null hypothesis can not be rejected.

A framework for aggregating sentiment measures in a mathematically coherent way is summarized in a road map.

Statistisk analys av språkliga sentimentmått

Sammanfattning

Företaget Gavagai använder olika mått för att i realtid uppskatta sentiment ifrån diverse strömmar av publika dokument. Gavagai vill hitta ett en procedur som bestämmer vilka mått som passar passar bäst i en given kontext. Det här arbetet diskuterar vilka kriterium som är önskvärda för att mäta sentiment samt härleder och utvärderar procedurer för att välja öptimalasentimentmått.

Tre metoder för att välja ut en grupp av mått som beskriver oberoende polariseringar i text föreslås. Dessa bygger på att: välja mått där principalkomponentsanalys uppvisar hög dimensionalitet hos måtten, välja mått som maximerar total uppskattad differentialentropi, välja ett mått som har hög villkorlig varians givet andra polariseringar.

Då exogen tidsvarierande data om ett ämne finns tillgängligt kan denna data användas för att beräkna vilka sentimentmått som bäst beskriver datan. För att undersöka potentialen i att välja sentimentmått på detta sätt testas hypoteserna att publika sentimentmått kan förutspå finansiell volatilitet samt politiska opinionsundersökningar. Nollhypotesen kan ej förkastas.

En sammanfattning för att på ett genomgående matematiskt koherent sätt aggregera sentiment läggs fram tillsammans med rekommendationer för framtida efterforskningar.

Contents

Acknowledgements

I wish to thank both my supervisors Jimmy and Jussi for their individual wisdom, help and interesting discussions. I also want to thank the colleagues at Gavagai, for helping on the sentiment extraction part, providing a vibrant atmosphere and discussing thoughts about natural language processing.

Word list

Sentiment — Feeling or emotion towards a specific topic.

Target — A topically coherent set of terms to represent some concept of interest in writing, such as a tradeable asset or a political issue.

Sentiment Variable — A time varying signal measuring a sentiment towards a specific target by aggregating information in documents that arrive at the stream during at a certain time point.

Polarization — A procedure to relate utterances or documents in a stream of text to sentiment measures of interest.

Pole — An exactly specified sentiment measure.

Document — A text that appears on the Internet and is analyzed, e.g. an article or a tweet.

Markovian Variable — A variable whose future only depends on its present value; it is independent of its history conditionally on its present value.

Weak stationarity — A stochastic process is considered (weakly) stationary if its mean and auto covariance function are constant in time.

1 Aggregating opinions from online text data

The information of *what every person expresses publicly* holds large potential value for prediction and decision making. Data describing this needs however to be presented concisely to be useful for a human to support decisions or make predictions.

In this work a data stream of different variables measuring sentiment are at hand and analyzed.

Previous work has been done on keyword-based sentiment analysis [17] [22]. Large scale sentiment analysis - or opinion mining - has become a possibility since then. This work will test methods to decide what subset of time varying sentiment variables extracted from public web documents by Gavagai that contain the most independent information. Information in this sense refers to what can be expressed visually for humans, to be able to draw conclusions about the target from.

Sentiment analysis trying to predict financial markets has been done with positive results [1] [11]. In this work financial prediction will be done to try to verify the predictive power of the sentiment analysis. It will also be used as a measure for selecting sentiment variables to present for a humans. Instead of predicting the price of an asset, the financial volatility of a market will be predicted.

Various techniques for visualizing multidimensional multivariate random variables have been studied and is summarized in [26]. Multidimensional refers to random variables with different physical dimension, which is the case of the variables output by Gavagai's sentiment analysis that will be used. Information entropy has been used in this context to explore similarity and dependency between random variables [5]. Principal Component Analysis has been used to visualize and simplify multivariate random variables [10]. In this work, techniques using information entropy will be used to measure similarity between a set of data points through parametric estimation. Also, Principal Component analysis will be used as foundation to quantify a measure of similarity between a set of sentiment variables, which should be as low as possible for variables containing independent information.

Anomaly detection for online data streams have been studied before through classification, clustering, nearest neighbor, statistical, information theoretic and spectral analysis among others. The comprehensive survey [2] compares these techniques and concludes that each choice of anomaly detection algorithm relies on a particular set of assumptions that each suit for a set of applications. In this work an anomaly detection algorithm is derived and implemented to detect and highlight the type of anomalies that are interesting for the specific application of sentiment analysis.

The prior hypothesis that online sentiment analysis has predictive power for financial markets will be investigated in this work, which will make the following contributions to the field of sentiment analysis:

1. Derive and implement an anomaly detection algorithm to highlight outlier events in the data stream.

2. Compare the use of: differential entropy, explained variance when adding an additional sentiment variable, and lost variance when excluding one principal component, to be able to select a measures of sentiment that contain a much independent information as possible.

3. Discuss possible improvements of the provided data stream that was analyzed; how the measured sentiment can be modeled mathematically and proposing an approach to use the information from online documents optimally by providing a road map for implementation.

4. Draw conclusions regarding the use of sentiment analysis for the specific contexts of political prediction, financial markets and business intelligence.

1.1 Goal and context

Gavagai is a text analysis company that uses language technology for the analysis of texts in real time and on Internet scale to provide actionable intelligence for market analysis purposes, political opinion tracking, and other related applications. One of the metrics Gavagai produce is a time line of general sentiment vis-a-vis some phenomenon, brand, political question or other target of interest, as expressed in social media, editorial media, or other publicly available texts.

Compared to other attempts to provide sentiment analysis Gavagai provides a broader palette of attitudes such as skepticism, desire, violence, or financial optimism. There are numerous parameters and methodological questions that have effect on the measurement of attitude in text, and Gavagai wishes to improve their current techniques to do this with greater reliability and validity.

Gavagai provides different types of intelligence, one of them is real time sentiment about a specific topic, or a target. The sentiment data that is studied takes the form specified in section 1.2 below. There are very many types of sentiment and ways of measuring them.

The problem that Gavagai wishes to solve is to: consider a set of sentiment time series data and select the most interesting sentiment variables to present visually for a client or analyst. The selection algorithm should work without any knowledge of the internal structure of the variables or how the measurements were made, i.e. a criterion based only on the raw data without the context of the data with the selection criterion to select the variables containing as much independent information as possible.

In essence; the problem formulation is to create a black box model that takes a set of time dependent input values and decides which subset of the variables are best to consider for a user.

1.2 Setup of data stream

We define a sentiment variable X_t for a time point t. It is is computed by aggregating documents that are published on the Internet during a time interval around the time point t concerning a specific target.f

Gavagai processes a total stream of documents $\mathcal{T}_t$ that consists of any document that is published on the Internet during a time interval around t. These documents are texts that are published as a news or blog article, in a discussion forum, publicly on Twitter or Facebook is counted. The time interval around t is chosen as either 1 hour or 24 hours depending on how often documents are published.

A target consists of a set of keywords K that are used to select web documents containing any of the keywords from the total stream of documents.

For any time point t there will be a set of documents D_t that contain at least one the keywords of the target during the time interval around the time point. This set reads as

$$D_t = \{d \in \mathcal{T}_t \mid \exists w \in K, w \in d\} \tag{1}$$

Each document $d \in D_t$ that match the target is analyzed for how much it exhibits the particular sentiment X_t which is quantified to a number $f_X(d)$. The measure of sentiment X_t is the aggregate of the documents for the time interval t. A sentiment variable X_t formed by summing over documents with a given measure $f_X(d)$ is referred to as a pole:

$$X_t = \sum_{d \in D_t} f_X(d) \tag{2}$$

A pole is hence uniquely described by the quantification of one document to the number $f_X(d)$. This quantification is done with a dictionary of words W that are likely to express the measured sentiment towards words that they co-occur with. A dictionary can be positive words for example, such as *great, happy, useful* etc.

The quantification to $f_X(d)$ is done by using the dictionary and checking if, or counting how many times, the targets keyword is in the proximity of any word in the dictionary.

Co-occurence, or the definition of proximity, is in general taken as words existing in the same phrase by Gavagai. Co-occurence is however not counted when a any word that can be used to negate a descriptive relationship between two words, such as *not, opposite*, etc.

Gavagai presents sentiment variables that can be either either a pole, or the

ratio of two different poles e.g. $Z_t = \frac{X_t}{Y_t}$

For example, the number of co-occurences between positive words and the target keyword is a sentiment variable, X_t. The number of co-occurences between positive words and the target keyword normalized by the total number of documents, $Z_t = \frac{X_t}{Y_t}$, is another sentiment variable.

The raw frequency, i.e. the total number of documents mentioning the target, is the pole when $f_X(d) = 1$. This pole can and will be considered interchangeably with sentiment variables.

The entire procedure is visualized in Figure 1.

The different poles have been shown to be useful ways of aggregating sentiment for Gavagai in different use cases. However, the total number of sentiment variables in the setup above are large and difficult to interpret manually without a quantitative measure of how much sentiment information they capture about a target discussed on the Internet.

1.3 What information is of interest to humans

A numerical representation of information taken from the web should preferably describe real concepts that humans are interested in. For example it is preferable that the units of measurements are easy to understand and have physical meaning.

What concepts are interesting to represent with sentiment analysis varies on the goal and application. In general any representation can provide value to a human analyst mainly in two different ways when covering a target:

- Real time-oriented - Discovering anomalies and sudden discussions that have arisen.
- Analysis-oriented - Comparing the sentiment from an underlying population that we want to measure. Comparing differences between targets or different time periods for one target.

The real time oriented information can in general be done by highlighting any unexpected events, e.g. with anomaly detection. The analysis oriented information often gives most information when relatively compared between different sentiment variables, and if possible comparisons between sentiment segmented over different demographics or document sources.

Specific use cases for sentiment analysis are discussed below.

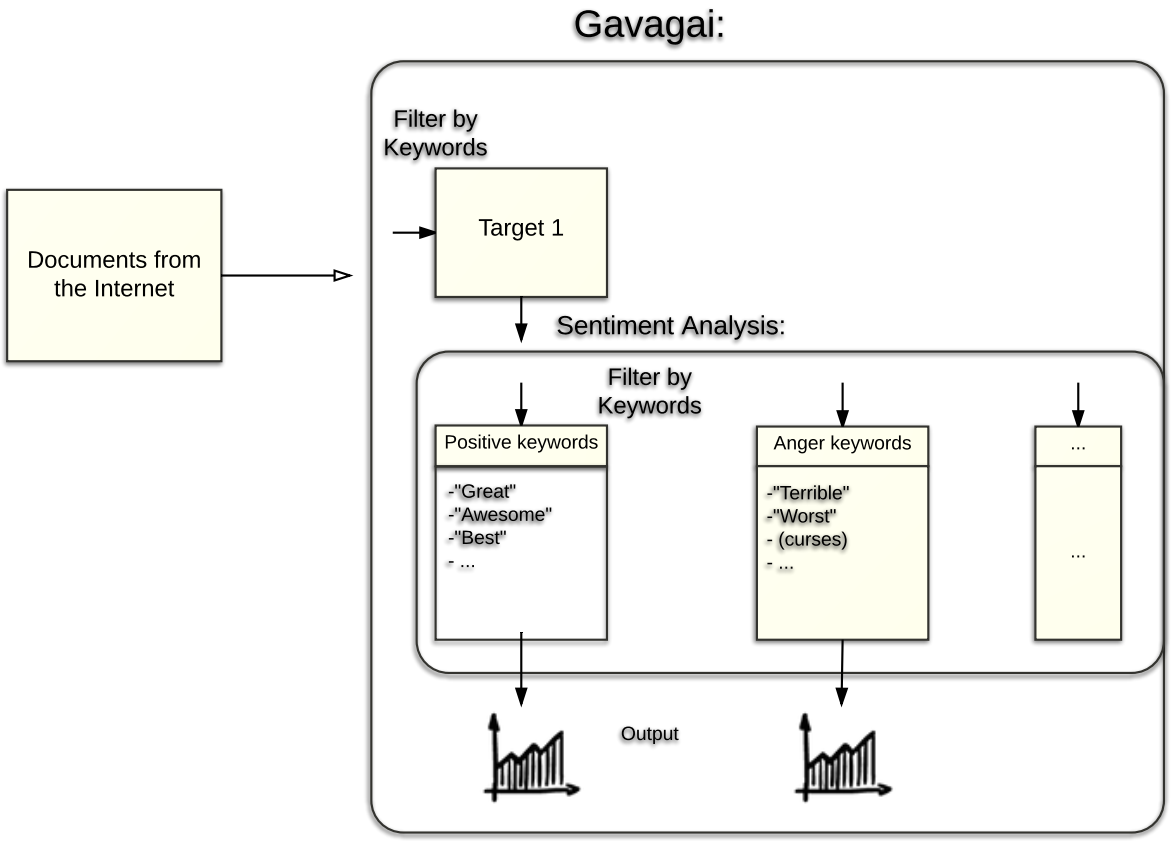


Figure 1: The data stream measuring sentiment is generated with sets of keywords as illustrated in the figure.

1.3.1 Three applications for sentiment analysis

The best model for analyzing sentiment variables is dependent on the use case. Below demands for three specific use cases are discussed with respect to technical possibilities. The demands for each case are partly based on interviews of people working in the three fields.

For analysts interested in financial sentiment, the information they tries to estimate might be: other investors confidence in the market, other investors implied willingness to invest and the sentiment regarding specific assets. They wish to aggregate sentiment over an as large portion of the population of investors opinions as possible, and if possible weight those by their influence, expertise or volume of capital to invest. Such a weighting can not be done directly with the sentiment variables at hand, but might be estimated by selecting a set of sentiment variables that contain this information. Segmenting analyzed documents by the source that published them also allows for estimating the influence of the measured sentiment.

For analysts interested in a marketing campaigns reach or a specific brands likability, they might want to measure the total number of mentions in social media over a longer time scale. Comparing the relative sentiment of mentions between different brands or over different time intervals would also be useful. Detecting sudden changes in how a brand is mentioned to be able to react is of interest. What targets are being discussed when a change occurs or the relative sentiment being used can be of interest.

For analyzing the political climate, detecting anomalies and changes in *how* a target is discussed is of interest. The sentiment of voters towards different political formations is also of interest. Representing the sentiment of the total population is difficult since different "types" of voters uses the Internet their opinions to a varying extent. For example, voters in the age 20-25 with radical political views possibly express their opinions on the Internet to a higher degree than voters aged 65 and over.

A possibility circumventing this is to segment the voting population and specifically track different types of voters separately, and measure the changes in their sentiment toward political formations. For such analysis to work well it would be necessary to keep track of authors of documents on the Internet, to decide to what demographic - or type of voter - they belong. Storing private information is something Gavagai actively avoids. However, the platform or domain used to publish a document to the Internet holds some information regarding the type or demographic of a voter. This could be see what type of voters opinion a specific document on the Internet represents. In summary, the sentiment towards political formations measured over all published documents is a biased sample from the sentiment of the voting population. The *changes* of the measured sentiment over time can however be used to represent the changes of the total voting population. Differentiating such changes from the noise from sentiment analysis is however difficult.

In all three use cases, segmenting documents by type of document and type of source, and aggregating the segments separately would be a useful to capture interesting information regarding the sentiment towards the topic. Estimating the number of individuals reached by documents or the influence of a document would also be of interest.

1.4 Choosing a model for analyzing sentiment variables

This section will consider what information in sentiment analysis is useful to a humans. Based on the conclusions a few possibilities for selecting sentiment variables semi-automatically will be suggested.

As introduced above, the real time-oriented and the analysis-oriented sentiment analysis are in general two different ways to use sentiment analysis. Selecting the sentiment variables that contain the most information is interesting to both of them. The semi-automatic selection suggested below selects variables for real time-oriented applications. Selecting sentiment variables for analysis-oriented applications needs to be be investigated qualitatively by using understanding of what is interesting to know about a topic.

Some part of this work will be devoted to discovering anomalies in sentiment variables. The discovery of an anomaly is in general done by putting assumptions on the behaviors of the measured time series and classifying a data point as an anomaly if it does not behave according to the assumptions with some a significance p. A simple such method for discovering anomalies will be derived and implemented below.

Many of the variables in the data streams will have lead-lag behavior, both internally for each variable and between variables. A possibility is therefore to fit a discrete time series model such as a multivariate Auto Regressive Moving Average model generated by Gaussian noise. The residuals, i.e. the estimate of the Gaussian noise, could then be consider the the real information [25]. Although this would be useful to make predictions and to compare the variables, it has drawbacks. Firstly, it processes the data while blind to the real meaning of the sentiment variables and knowledge regarding how they relate to each other. Secondly, it puts strong assumptions on the regularity of the time series. Sentiment variables are generated from very complicated phenomena where stationary dependence between measurements in time and Gaussian distributions are difficult to motivate.

A potential solution to these drawbacks is to learn a state-space model with parameters that describe how the measured sentiment variables are generated. This model could account for a change in how documents are published and measured.

After inference on the data the state-space would contain information such as the the estimated intensity of new documents for a target, and ratios of sentiment in published documents. These parameters are quantities with physical meaning

that we are interested in. An example of this follows below.

Example 1.1. Assume that the number of documents matching a target equals D_t for a time interval t for $t = 1, 2....$ D_t is distributed according to $D_t \sim Po(\lambda^0 + \lambda_t^{\text{trending}})$.

Here λ^0 represents the stationary intensity of published documents over the interval, and $\lambda_t^{\text{trending}}$ represents the additional trending intensity. Moreover, let the probability that a document is classified as having positive sentiment towards the target be p^0 for documents that appear with stationary intensity λ^0, and p_t^{trending} for documents that appear with a trending intensity $\lambda_t^{\text{trending}}$.

Then assume that λ_t and p_t^{trending} are Markovian, and more specifically change pairwise in the two cases:

- $\lambda_t^{\text{trending}}$ is partly remembered, it decays with a constant and noise is added, and p_t^{trending} is the same as the interval before
- $\lambda_t^{\text{trending}}$ is set to a completely new value drawn from the distribution ϵ_t and p_t^{trending} is drawn from a Beta distribution.

This model captures that there are discrete breakpoints where new types of discussions are started regarding the target. The intensity associated with a new discussion will either increase or decrease from its previous value, or be replaced by a new discussion (which is allowed to be the special case of a discussion with zero intensity). The probability that each new document associated with a discussion has a certain sentiment is assumed to have a constant sentiment p_t^{trending}, until the next discrete breakpoint.

The first case happens when $b_t = 1$ and has a constant probability q_{increase}. The proposed choice for the change of the intensity ϵ_t is an Exponential distribution (motivated by that it is the continuous limit of the geometric distribution).

$$\begin{aligned} \lambda_t &= \begin{cases} c\ \lambda_{t-1} + \epsilon_t & \text{if } b_t = 1 \\ \epsilon_t & \text{if } b_t = 0 \end{cases} \\ \epsilon_t &\sim Exp(\mu) \\ p_{pos,t} &= \begin{cases} p_{pos,t} & \text{if } b_t = 1 \\ Beta(\alpha, \beta) & \text{if } b_t = 0 \end{cases} \\ b_t &\sim Be(q_{\text{increase}}) \end{aligned} \tag{3}$$

Choosing priors for the parameters is then possible. In the particular case of conjugate priors this would be:

$$
\begin{aligned}
\mu &\sim \text{Gamma}(k, \theta) \\
\lambda^0 &\sim \text{Gamma}(k', \theta') \\
p^0_{pos} &\sim Beta(\alpha, \beta) \\
p_{\text{increase}} &\sim Beta(\alpha', \beta') \\
c &\sim \text{Gamma}(k'', \theta')
\end{aligned} \tag{4}
$$

Where the prior for $p_{pos,0}$ and $p_{pos,t}$ are the same.

Finally, by counting the number of positively classified documents P_t and the total number of documents D_t for a target it is possible to form the distribution of the model parameters conditionally on the data through the relation below.

$$
\begin{aligned}
D_t &= D^0_t + D^{\text{trending}}_t \\
D^0_t &\sim Po(\lambda^0) \\
D^{\text{trending}}_t \sim Po(\lambda^{\text{trending}}_t) P_t &= P^0 + P^{\text{trending}}_t \\
P^0_t &\sim \text{bin}(D^0_t, p^0) \\
P^{\text{trending}}_t &\sim \text{bin}(D^{\text{trending}}_t, p^{\text{trending}}_t)
\end{aligned} \tag{5}
$$

The parameters posterior distribution can then be drawn with MCMC methods such as the Metropolis-Hastings sampler or by using sequential Monte Carlo methods, to reduce computational demand despite large amounts of data, if necessary.

□

Using a state-space model as in the example above could provide knowledge regarding which sentiment variables measure trending changes for a target and which measure stationary sentiment. The estimated distribution of the parameters could be used to measure how sentiment has varied in time, and how much is systematic randomness compared to how much is an actual change in sentiment of individuals. With a more sophisticated model the interdependence between sentiment measurements could be taken into account exactly.

Such models, which may vary between different use cases, however require a lot of tuning. The goal in this work is to attain a simple black box model for any type of sentiment variables, so such a model will not be implemented. Furthermore, strong assumptions on the data stream at hand are dangerous since the process generating the sentiment is changing over time. For example what type of documents that are usually published, if particular targets are being discussed with the sentiment expressing keywords or in more subtle texts, and how different syntactic expressions are trending on the Internet.

A state-space model for sentiment analysis with the data stream at hand will not be tested practically in this work. However, a similar approach for aggregating

the entire distribution for the measured sentiment is suggested in the road map for efficient implementation of sentiment analysis that is presented in the end of this work.

The methods that will be tested in this thesis for drawing conclusions about the data stream are introduced below in section 2. In summary the methods will be as follow:

The first method will try to predict exogenous data with different sentiment variables with a simple model through regression. It will also let us test the sentiment variables for causality to the exogenous data with the Granger Causality test.

The second method will measure unpredictability of sentiment variables, with respect to each other, and interpret this as the independent information for a chosen set of sentiment variables over a target. This will be done with three approaches by:

1. Measuring the lost variance when excluding one Principle component from the data.
2. Measuring unpredictability as the estimated differential entropy of the multivariate data set.
3. Measuring the linearly unexplained variance when expanding a set of sentiment variables with another sentiment variable.

A method for detecting anomalies in the data stream will also be derived and implemented and used in the other methods.

Practical implications from the mathematical point of view regarding the sentiment analysis data stream will also be discussed, and conclusions drawn, in the ending section.

2 Mathematical background

2.1 Principal component analysis

Principal component analysis [20] (PCA) is a statistical procedure that performs an orthogonal transformation to convert a set of observations of correlated random variables to a basis where each base vector is uncorrelated. The number base vectors, or principal components, is less or equal to the number of original variables. The transformation is defined so that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (and thereby uncorrelated with) the preceding components.

Principal component analysis has been used in many applications in science with multivariate data for exploratory analysis, and can be thought of to find the inner structure that explains most of the variance.

Maximizing the variance of the first component of any observation matrix X that has been centered around its mean is equivalent to solving:

$$\begin{aligned} &\underset{w}{\text{maximize}}\ \text{Var}(Xw) = (Xw)'Xw, \quad |w| = 1 \\ &\text{Lagrange relaxation} \Rightarrow \\ &\frac{d}{dw_i}(Xw)'Xw + \lambda(|w|^2 - 1) = 0, \quad |w| = 1 \\ &\Rightarrow X'Xw = -\lambda w \end{aligned} \tag{6}$$

Which gives that w is an eigenvector of the sample covariance matrix $X'X$, which is symmetric. Since all the eigenvectors of a symmetric matrix are orthogonal, they are equal to the principal components u_i. Furthermore we have that the variance is equal to each principal components eigenvalue:

$\text{Var}(Xw) = (Xw)'Xw = wX'Xw = \lambda w'w = \lambda$

PCA is sensitive to the relative scaling of the original variables, hence observations of variables with different dimensions should be scaled to avoid ambiguity. Scaling them to unit variance is a good choice if the variables are considered equally important.

The first $N-1$ principal components must account for between $\frac{N-1}{N}$ up to 1 fraction of the total total variance of the data. The case when they account for all of the variance is when the observations are all linearly dependent, i.e. they lie on a hyperplane of dimension $\leq N-1$. In this case the data can be said to contain low amount of information.

The idea is to use PCA on sentiment data and compare how much variance that would be lost from the observed data if the principal component with the lowest variance was excluded.

2.2 Differential Entropy as a measure of information content.

Differential entropy is a generalization of the discrete case Shannon entropy. It has been studied in information theory for various applications and is a measure of unpredictability of a random variable [24].

For a random variable with probability density function f the differential entropy is defined as follows:

Definition 1 (Entropy)**.** $h(X) = -\int_{\mathbb{X}} f(x)\log f(x)\,dx$

The differential entropy of a continuous random variable can be estimated from a sample of the variable. This can be done by dividing it into discrete histograms [6], through kernel density estimation [9] or fitting it to a parameterized distribution that has a closed form for the differential entropy. One such distribution is the multivariate Gaussian distribution, which has the entropy [14]:

$$\frac{k}{2}(1 + \ln(2\pi)) + \frac{1}{2}\ln|\mathbf{\Sigma}| \tag{7}$$

Where k is the number of dimensions and $\mathbf{\Sigma}$ is the covariance of the Random Variable.

By fitting the Gaussian distribution to any dataset the differential entropy can be used to identify non-linear and complex covariation in sets of random variables [6].

To decide which of two subsets of random variables with samples contains the most independent information, a possible approach is to compare the estimated differential entropy of the two subsets.

2.3 Stationarity and Causality

The Granger Causality is a simple way to test the null hypothesis of non-causality for the time series Y dependence on X. This is done by fitting a linear model between lagged values of X to predict Y, and testing the significance of the fit.

An important assumption on the time series is first needed, which is that the time series X and Y are both stationary [7]:

Definition 2 (Weak stationary)**.** A discrete stochastic process X_t is considered stationary if $E(X_t)$ and $Cov(X_{t+h}, X_t)$ are independent of t for each h.

The discrete time series Y is modeled as an autoregressive moving average time series. This refers to a time series where each data point is a linear combination of previous values plus a white noise error.

$$Y_t = \sum_{i=1}^{p} \varphi_i Y_{t-i} + \varepsilon_t \tag{8}$$

Stationarity for an autoregressive moving average time series can be tested through the Augmented Dickey Fuller test. The null hypothesis of the Augmented Dickey Fuller test is that there is a unit root of the characteristic polynomial. The characteristic equation is one minus the polynomial with the i:th coefficient equal to φ_i above. If there is no unit root it means that the time

series is stationary. The Augmented Dickey Fuller test rejects the null hypothesis if the t-distributed test statistic falls below the critical value for the sample. The test is implemented in most statistical software packages.

In the Granger Causality test, k lagged values of Y are used to predict itself with a linear model which is the multiple linear regression, which is discussed below. The purpose of this auto regression can intuitively be seen as to remove information in X about Y that is already present in Y itself.

After deciding the number of lagged parameters the test for significance of the unrestricted model is tested against the restricted model. The unrestricted model uses all lagged variables to to predict Y, whereas the restricted model uses only lagged values of Y itself to predict auto regressively.

Comparing these two models makes it possible to form the F-statistic to get the significance of either rejecting the null hypothesis. The F-statistic is formed as follows:

$$F = \frac{RSS_R - RSS_{UR}/m}{RSS_{UR}/(n-k)} \tag{9}$$

Where RSS_R is the residual sum of squares of the restricted fit and RSS_R is for the unrestricted fit. m is the number of parameters in the restricted model, n is the total number of data points and k is the total number of parameters in the unrestricted model.

The expression for the F-statistic is that of the ratio of the normalized explained variance over the normalized variance without he predictive variable. Under the null hypothesis this belongs to a F(m, n-k)-distribution, the distribution of a $\chi^2(m)$ random variable over a $\chi^2(n-k)$ random variable, where m and $n-k$ degrees of freedom. The null hypothesis, that X does not predict Y, is then tested against that X does predict Y by computing the critical value for the F-distribution for a given significance level α.

Important assumptions of the regression is that the residuals are independent and normally distributed. Normality is in general efficiently inspected by plotting the empirical vs the theoretical quantiles of the distributions. That they are not correlated, which is equivalent of being independent if the variables do belong to a multivariate normal distribution, can be tested with the Ljung box Q-test [25].

2.4 Multiple Linear Regression and Regularization

Linear regression considers a model where the conditional mean of y given the value of X is an affine function, or a first order Taylor approximation, of X: $E(y) = \bar{a} \cdot \bar{x} + b$. The assumption that the deviation from the mean is independent and identically normally distributed noise (referred to as residuals), gives the

maximum likelihood estimation of a as equal to the Ordinary Least Square fit of a from the data points in $\bar{y}$ X:

$$\bar{a} = (XX^T)^{-1}X\bar{y}$$

The assumption of independent normally distributed residuals also lets us compute the distributions and confidence intervals of a and the sum of squared errors of the fit [13]. This gives methods to test the validity of the regression.

A useful measure of the fit is the fraction of explained variance of the fit:

$$R2 = \frac{Var(Y - \bar{a} \cdot \bar{X})}{Var(Y)} = \frac{SS_{\text{tot}} - SS_{\text{res}}}{SS_{\text{tot}}} = 1 - \frac{SS_{\text{res}}/n}{SS_{\text{tot}}/n}$$

where SS_{tot} is the sum of squares of the dataset and SS_{res} is the sum of squares of the residuals, which are uncorrelated under assumptions.

If we consider the variances above to be estimates and replace them with their unbiased estimates the *adjusted-R2* $\bar{R}^2$ is attained, which is more suitable for selecting model as it takes into account the degrees of freedom of the mode [23], and we get:

$$\bar{R}^2 = 1 - \frac{SS_{res}/df_e}{SS_{tot}/df_t}$$

where df_t is the $n - 1$ degrees of freedom of the estimate of the total variance, and df_e is the $n - p - 1$ degrees of freedom of the estimate of the residuals of the fitted model.

To avoid over fitting with many parameters it is useful to assume a prior distribution on the parameters so that the model fit prefers absolute values of parameters to be small. This is a form of regularization, which makes any regression more robust to noise in data and spurious correlation. Two types of priors used are preferable: a zero centered Normally distribution or a Laplace distribution, yielding Ridge Regression and Lasso Regression respectively [18].

To decide the priors for the parameters in the regression the dataset can be k-fold cross-validated. This just means to repeatedly evaluate the regression on a part of the dataset that was not used for regression, and choose priors so that the regression generalizes well even to unseen datasets.

2.5 Cross validation

For any model that fits parameters to data to minimize a cost function such as the negative likelihood or squared errors, it is useful to know how good the model fits for data from the same population that was not used to fit the parameters. This is done with cross validation . [12]

In the typical setting k-fold cross validation is used. This means that the entire dataset is split into k equally sized subsamples. The first $k - 1$ sample are then used as training data to fit the model, and the last unseen sample is used

to evaluate the fit. This is then repeated k times with each of the samples used as validation set once. The results of the evaluations are then averaged (or otherwise combined), to produce a measure of model fit without evaluating data that was used to fit the model.

Cross Validation can and should be used on any type of model to check that the model is not to complex but generalizes well to unseen data.

2.6 Bootstrapping

When computing a metric or performing a test on data, it is often useful to estimate the uncertainty of the metric or test. This can be done by drawing new samples from the empirical distribution of the data, and compute the metric each times to attain the *bootstrapped* distribution of the metric. [3]

To draw samples from the empirical distribution with replacement and evaluating statistics on these samples is referred to as bootstrapping. The strength of the method is that it can be done for any metric without assumptions of the assumptions, and yield estimates of confidence intervals or significance of a metric or statistical test.

3 Analyzing and selecting sentiment variables

In this section we describe how the methods introduced above are implemented.

3.1 Dependence on time and weekday

The number of published documents varies throughout the day as well as the week and so does the measured sentiment variables. It is important to have this seasonality in mind when performing regression with the variables.

Two parameterizations are suggested for de-seasoning. These are tested for the targets Global Economy and Bitcoin that are defined and measured by Gavagai.

De-seasoning, 1. For a sentiment variable we denote the measured sentiment for every timepoint t by U_t. We let X_t be the de-seasonalised sentiment variable.

$$U_t = X_t + avg(U_{weekday(t),hour(t)})$$

Where the indexes $weekday(t)$ and $hour(t)$ refer to the weekday and the hour of the time-point t.

De-seasoning, 2. As an alternative to the above, we measure the daily average and the hourly average above the mean separately. These can be used to subtract the current hours average as well as the current weekdays average simultaneously.

$$U_t = X_t + avg(U) + (avg(U_{hour(t)}) - avg(U)) + (avg(U_{weekday(t)}) - avg(U))$$

The latter method is introduced as it is preferred when we have too few data points to yield smooth averages for every combination of the 7 weekdays and 24 hours.

3.2 Predicting the SP500 volatility index

We test the hypothesis that online sentiment variables has predictive power of financial markets with the Granger Causality test. More specifically we try to predict the 30 days implied volatility of the S&P500 as measured by CBOE VIX.

The explained variance by the prediction amount of causality can also be useful in selecting or designing a sentiment measure capturing as much information as possible.

Dataset. The VIX index was taken between 2014-07-30 - 2015-01-25 from CBOE Exchange. We let the predicted variable Y be the daily log-returns of the VIX Index. The log-returns are computed as the logarithm of the ratio of subsequent prices:

$$Y_i = \ln(P_i/P_{i-1})$$

The quoted opening as well as the closing price were used computing the nightly and daily log-returns - so that the average sentiment X during the night can be used to predict the days return and the sentiment during the day can predict the nights return without any overlap.

The sentiment variables evaluated were the sentiment poles Positivity, Negativity, Worry, Volatility, FinancialUp, FinancialIncrease and FinancialDown for the target Global finance for the same time period as the prices. The sentiment variables were checked for not containing outliers with the model as described in 3.3.

Preprocessing. The seasonality of the sentiment was inspected to be similar through weekdays but not weekends and or mornings before Mondays, see Appendix. The daily variation for weekdays was similar over the day. Sentiment was thereby taken as the mean during three separate types of intervals; Weekends, nights between two trading days and days of a trading day. The sentiment mean during these intervals was de-seasonalised by removing its mean, as well as normalizing by variance.

Stationarity. To asses the assumption of stationarity, the Augmented Dickey-

Fuller test was performed testing the null hypothesis of a unit root, i.e. non-stationarity, for both the predictor and the predicted variable.

Method. The dependent variable Y is assumed to be a linear function of the sentiment variables measured earlier in time and previous lags of Y itself. Since more than one lag can be used it makes it possible for the model to to capture numerical approximations of derivate of the variables.

The BIC statistic of the model fitting the Y to itself was first minimized by choosing the number of lags in the auto regression. The number of lags of X to be used together with the lags of Y was then decided by minimizing the BIC statistic of the entire model. The F-statistic for adding the exogenous parameters was calculated and the the null hypothesis of non-causality was tested versus causality as described in section 2.3.

Non-autocorrelation of the residuals was tested with the Ljung box Q-test. The residuals were also plotted against time to asses independence of the residuals. To check for equal variance of the residuals, they were plotted against the fitted values of the log-returns and checked for any non-random pattern.

The assumption of normality of the residuals was checked with empirical vs theoretical quantile-quantile plots.

The adjusted R2 with the autoregressive model as benchmark was used as a measure of estimated explained variance by the sentiment variables. How well the entire model generalized was also tested with 10-fold cross validation.

3.3 Anomalies

It is desirable to identify when something has changed about how people speak about a target. A simple approach identifying this derived from how the sentiment variables might change is evaluated below.

Model selection.

There are very many possible choices when automatically classifying data points as outliers, as discussed above and in [2]. The best model will in most cases leverage assumptions that can be made for the specific application. Assumptions can make it possible to have a simple and straightforward model, which is advantageous since it is more often robust to unforeseen changes of the data, and can be applied in many different contexts.

The aspects for this application that are leveraged for choice of model are as follows:

1. A decrease in a sentiment variable is in general not an interesting anomaly, compared to a sudden increase. The natural argument for this is that

if documents published on line suddenly contain less keywords it is not interesting to bring attention to rather than the opposite.

2. Sentiment variables are in general proportional to the raw frequency of the mentioned target.

3. Most of the anomalies will be discovered by considering just the raw frequency of detected documents. The rest of the interesting anomalies - in how a target is discussed - will strongly impact the ratio of sentiment variables and the total number of scanned documents.

Assumptions. From the specifications above we normalize every sentiment variable by the raw frequency F_t to get X_t. We say that each sentiment variable X_t as well as the raw frequency F_t is then normally distributed for every data point where it is *not* an anomaly. Data points that are anomalies are assumed not to lie further away than the $p = 0.01$ confidence interval of the normal distributions.

Method. We find anomalies in two steps, we first remove obvious outliers to get a better estimation for the measured distributions as assumed above, and then find anomalies with better estimations of the normal distribution parameters. On average 1 data point will be misclassified as an anomaly in the first step if there are no anomalies in reality. The both steps are basically identical but the parameter of trigger significance, or the average misclassification rate when there is only natural variation from the distributions assumed above, can be chosen in the second step. The sentiment variables and the raw frequency is treated equally.

A one sided interval (λ_X, ∞) for each sentiment variable X_t that decides if the measurement should be classified as an anomaly or not is formed from estimated parameters. Since the variance is estimated and no known the confidence interval is taken as the $1 - p$ quantile of the Students t distribution with N degrees of freedom, where N is the number of data points. The significance p_i is set as $p_i = 1/N/k$, k being the number of sentiment variables, so that 1 data point is classified as an anomaly on average over the entire dataset with the above assumptions even if there are no anomalies (see details below).

To improve the estimation for the mean in this initial outlier-detection the median is used instead of the mean to make the estimation more robust to outliers caused by X_t not being normally distributed. The median being an unbiased estimate of the mean for normal distribution.

The intervals bound λ_X is then computed from the Students t distribution and all anomalies that are outside the intervals are classified as anomalies and removed.

The process is repeated with a new interval λ'_X but this time with better estimates of the mean and variance. The significance is then set to p_i which can be chosen manually, or set to $p_i = 1/N/k$ to only lose on average 1 other data point - if there are no anomalies.

Confidence level. When triggering on any of many possible variables the aggregated significance level p_{tot} of a anomaly classification becomes larger. Since the variables can be correlated even after the transformation above, we do not know what the actual confidence level is.

If we assume that the triggers are independent the upper limit of the aggregated trigger significance p_{tot} is reached as:

$$p_{tot} = 1 - (1 - p_i)^k \tag{10}$$

where p_{tot} is the aggregated significance level and k is the number of independent used triggers.

In practice attaining a desirable significance p_{tot} can be done by choosing it manually to then compute the individual significance for each sentiment variable from equation 10, and then adjusting them if the trigger rate seem higher or lower than desired, i.e. if the significance was chosen too high or low.

The choice of letting the significance $p_i = 1/N/k$ to mis-classify on average $1/N$ outlier per data point above, when there are no anomalies, is motivated by taking the assumed aggregated significance:

$$p_{tot} = 1 - (1 - 1/N/k)^k$$

And assuming that $p_i = 1/N/k$ is small to get $p_{tot} = 1/N$, as seen when expanding:

$$1 - (1 - p)^k \; 1 - (1)^k - k(1)^{k-1}p + O(p^2)) \approx 1 - (1 - kp) = kp = k/N/k = 1/N$$

3.4 Analyzing similarity in a set of sentiment variables

Principal Component Analysis and differential entropy can be used to measure the similarity or unpredictability of a set of sentiment variables. Projection through linear regression can be used to measure how much unexplained variance is added with a new sentiment variable. We evaluate these metrics as a measure for the independent information of sentiment variables.

Dataset. For all the three methods evaluated, the information contained in sentiment variables for the target Mcdonalds was analyzed. The measures of information content, was compared for the sentiment variables Positivity, Negativity and Trust are compared to the variables Positivity, Negativity and Skepticism.

Visualization and Preprocessing. The first step in selecting sentiment variables is to inspect the variables for correlation and/or nonlinear relationship.

This was presented in scatter plots and inspected. Since the different sentiment variables have different orders of magnitude it is possible to normalize them by variance if they are considered equally important. Since the variables all measure a sum of word co occurrence counts, normalizing loses the original dimension and the correct relationship between variables. Both original and normalized variables are tested but the normalized variables are used for all the evaluations.

Before normalizing, outliers are removed as described above in section 3.3, and the variables are de-seasonalised on a weekday basis as described above in section 3.1.

Principal Component Analysis. For a given set of sentiment variables we assume that they were pairwise correlated to a number that is constant in time. We then carry out the Principal component Analysis algorithm to get the principal components, i.e. the uncorrelated linear combinations of the sentiment variables and, more importantly, the relative explained variance of each component.

If a large amount of the variance is explained by the first few Principal Components it means that the whole set of sentiment variables are varying strongly together, i.e. can be assumed to measure the same information. Therefore the lost variance when excluding one principal component was computed, for the two sets of sentiment variables.

Differential Entropy. Assuming that the set of sentiment variables is a multivariate Gaussian distribution, the differential entropy can be computed as in section 7. This was done for two sets of sentiment variables, with one variable differing, to test. Even if the normality assumption does not hold particularly well, it is assumed that estimated entropy gives a measure of the normalized multivariate distributions unpredictability which is needed.

Projection. We considered an initial set of sentiment variables and estimate how much additional information some other sentiment variable would add.

This is done by assuming a linear relationship with the new variable and performing the ordinary least squared fit to gain the parameters, and measure how much variance that was still left in the residuals. The unbiased estimation of the explained variance of the fit is the adjusted R2, and hence it was used to calculate unexplained variance.

Bootstrapped confidence intervals. To test if the difference of the independent information measuring variables for the two sets was significant, confidence intervals were estimated for them. This was done by performing 10^4 bootstrap samples over the dataset and evaluating the difference between the two sentiment variables for each sample. The confidence intervals were constructed by using the Bias-Corrected Accelerated Non-Parametric Bootstrap [3] which adjusts for skewness and bias. Significance was tested by checking if the null hypothesis, i.e. that the difference was equal to zero, could be rejected which is equivalent to zero being outside the confidence interval.

3.5 Regression to price data

Sentiment variables are selected for the target Bitcoin by visualizing the correlation between all of the sentiment variables as well as the log-return of the bitcoin price.

Regression is performed to see if sentiment variables do hold predictive value to the price of bitcoin.

Price data. The trades of Bitcoin at Bitstamp [4] was used to calculate the daily volume weighted average price. Sentiment variables listed below and prices between from the 1 August 2014 and 100 days forward was used. Log-returns are defined as the natural logarithm of the ratio of two consecutive volume weighted prices.

Preprocessing and Visualization. For the regression the outliers were discarded with the anomaly detection as with the model as described above. The data was de-seasonalised on a weekday basis, and the variables were normalized by their mean and standard deviation.

The pairwise correlation between each sentiment variables and the log-return was plotted in color to give overview of how they were related. This could in practice be useful to choose a set of sentiment variables manually.

Method. We consider one sentiment variable at a time and an intercept to be linearly related to the price change of Bitcoins. The ordinary least squares minimization is then used to compute the parameters of the model.

The explained variance and cross validated explained variance was computed as a measure of the predictive value of the sentiment variable. The assumptions of linear regression was not used, instead the average 50-50 folded cross validation was evaluated with and without regularization as described below to measure the generalization of the model.

Regularization. The linear model was regularized with the lasso regularization. The parameters for the regularizations were found by performing multiple 50-50 folded cross validated splits and optimizing to maximize the R2 coefficient of determination over the test set to achieve good generalization. This method makes it possible to only compare the R2 value between different sentiment variables without relying on the model assumptions of the linear regression into account to estimate the predictive power of different sentiment variables.

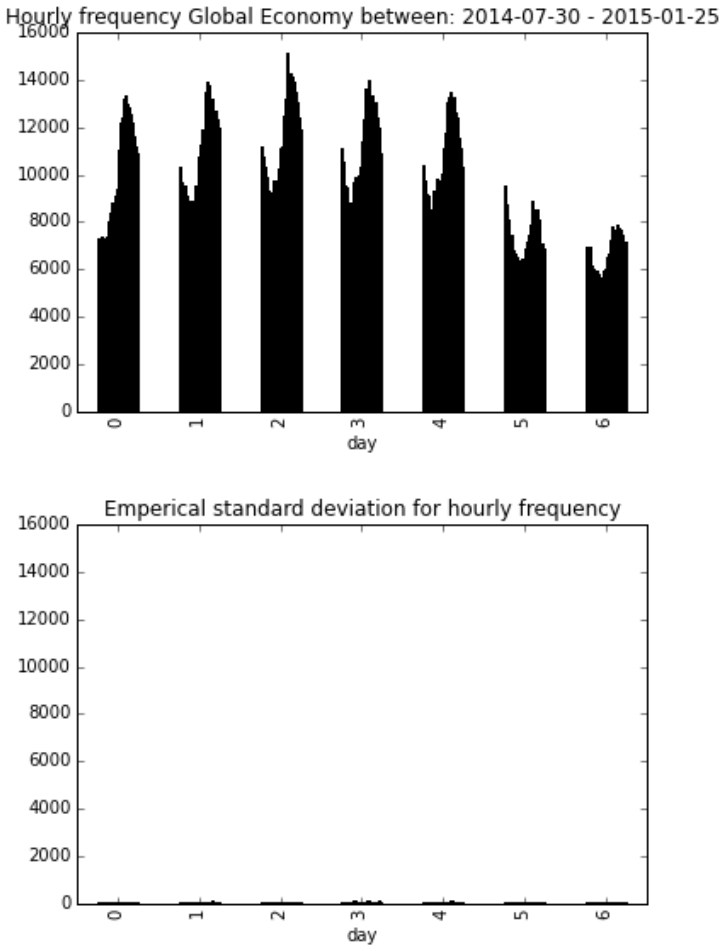


Figure 2: Hourly mention of Global Economy and empirical standard deviation Monday - Sunday, starting 00:00 CET.

Sentiment Name:	adj R2:	Crossval R2:	p:
Positivity	0.003	0.005	0.368
Negativity	0.003	-0.017	0.419
Worry	0.003	-0.070	0.417
FinancialVolatility	0.003	-0.029	0.436
FinancialUp	0.002	0.004	0.499
FinancialIncrease	0.010	-0.025	0.126
FinancialDown	0.004	-0.016	0.314

Table 1: Adjusted R2 (estimated variance explained), cross validated R2 and significance p for each sentiment when testing for Granger Causality to the VIX, with a 3 lag autoregressive model as benchmark.

4 Results

4.1 Dependence on time and weekday

Seasonality was inspected for hour resolution data, total number of mentions for the Global Finance target is shown weekly and daily for in Figure 12.

There is a clear hourly seasonality while the daily distribution is roughly equal for weekdays, and a bit lower for weekends.

4.2 Predicting the SP500 Volatility Index

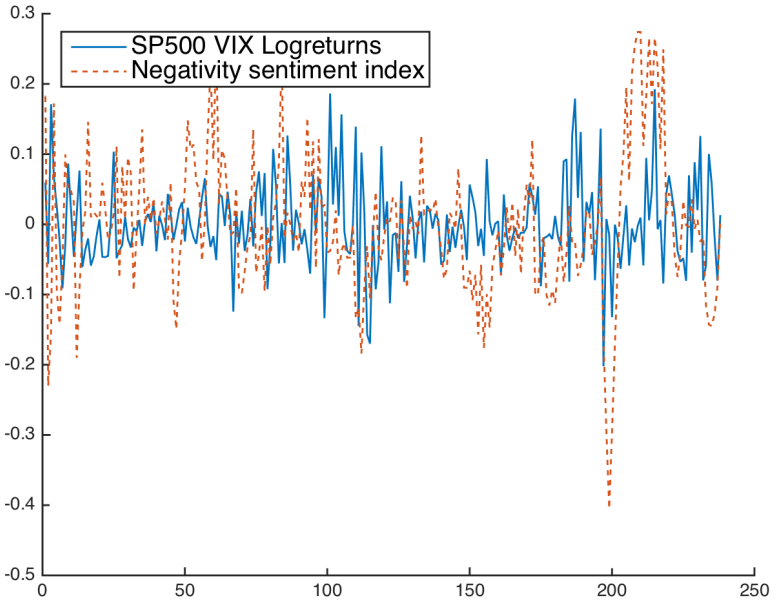
The test for causality described in section 2.3 was performed on the SP500 VIX data and sentiment variables, listed in Table 1.

Minimizing the BIC lead to a model with 3 lags of auto regression and 1 lag of exogenous sentiment variable values.

The Augmented Dickey-Fuller test showed that null hypothesis of non-stationarity could be rejected for Y as well as each sentiment measure, with significance $p < 10^{-3}$.

The Box-Jenkins test showed that no-autocorrelation for the residuals in the fit could not be rejected i.e. that the residuals seemed to be uncorrelated, for each sentiment variable. This is visualized in the Appendix.

The quantile-quantile plots in Figure 4 show that the distributions partly deviate from the assumption of normality; the distributions partly exhibit fat tails. The consequence of this is that the true p-value of the Granger test slightly differ from what is calculated above.



(a) VIX log-returns and the sentiment measure for Negativity one lag behind, sampled twice per day.

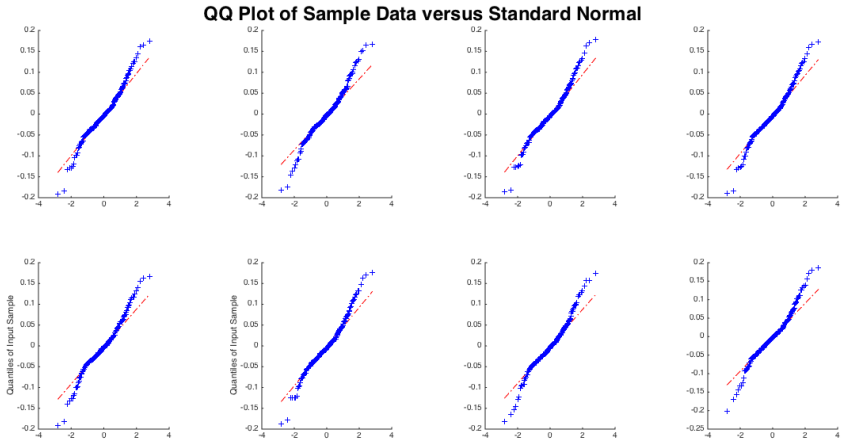


Figure 4: Empirical quantiles vs normal distributed quantiles of each residual for the regression to the VIX with sentiment enumerated as in the table above. The dependent variable itself in the last subplot.

Figure 5: Anomalies for the target McDonalds, shown in in dotted lines. The detection triggers on any sentiment variable. The first plot contains all data points, the anomalies are then removed iteratively for the two plots below.

4.3 Anomalies

The anomaly detection method described in section 3.3 is used and triggered days are shown with red bands in Figure 5.

The method triggers when there is a fluctuation in the sentiment variables of less magnitude, which is difficult to perceive from the figure as it is small relative to the other larger variables.

4.4 Analyzing similarity in a set of sentiment variables

Scatter plots for each sentiment together with the raw frequency for the target McDonalds can be seen in Figure 6 both in their original dimension and normalized. The total cumulative variance for each principal component for the two sets of sentiment variables can be seen in Figure 7. The final de-seasonalised sentiment variables used can be seen in Figure 8.

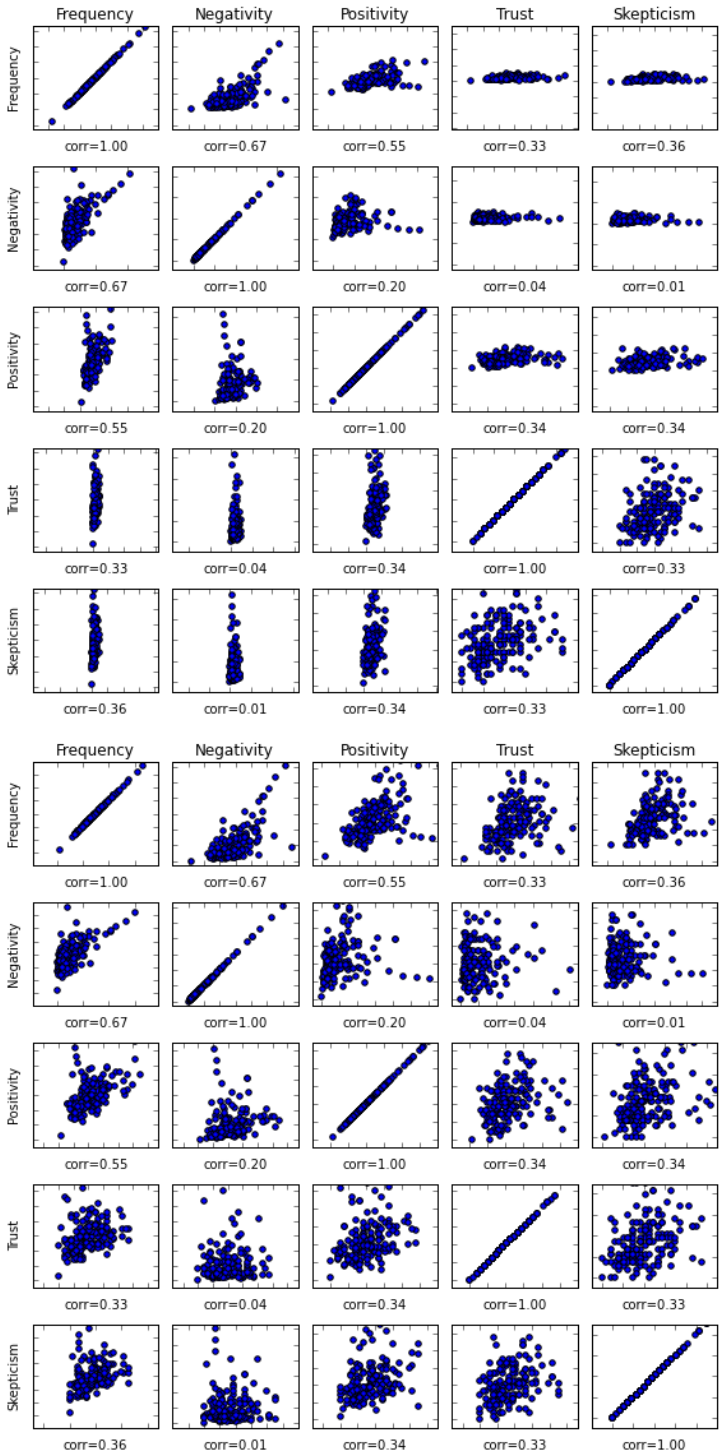


Figure 6: Pairwise scatterplots between sentiment variables and the correlation. Data for the target McDonalds. Below the variables are scaled to unit variance, the two variants are compared.

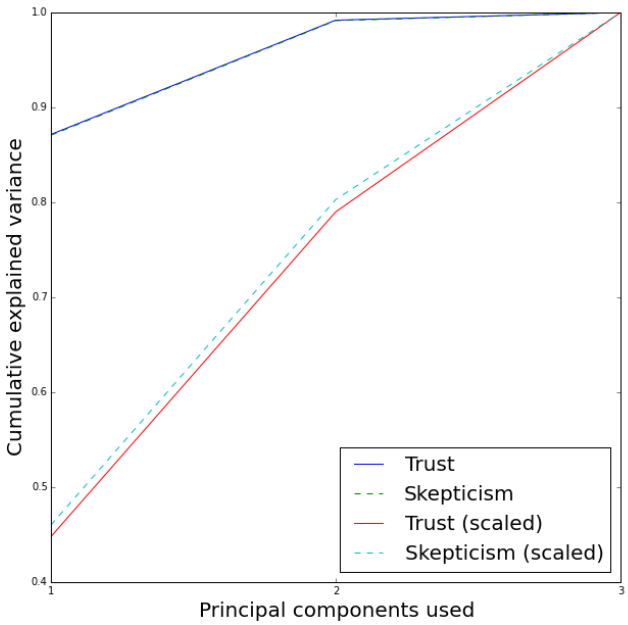


Figure 7: Cumulative explained variance for the first principal components, for the sets Positive, Negative, Trust or Positive, Negative, Skepticism shown for the original and nomralized sentiment variables.

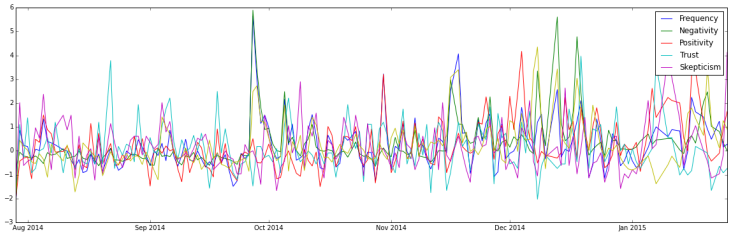


Figure 8: The scaled and de-trended sentiment variable McDonald that was used when measuring independent information in sentiment variables.

Measure of added information:	Trust sentiment	Skepticism sentiment
Lost variance when excluding one PCA component	0.210	0.197
Estimated entropy	4.195	4.178
Unexplained variance in last variable	0.894	0.863

Table 2: Measures of independent information when using the set of sentiment variables Positive, Negative, Trust compared to Positive, Negative, Skepticism. Data is from the target McDonalds.

Measure of added information:	Difference	Bootstrapped confidence intervals
Lost variance, excluding one PCA component	-0.0130	(-0.0621, 0.0331)
Estimated entropy	-0.0173	(-0.230, 0.181)
Unexplained variance in last variable	-0.0304	(-0.165, 0.0912)

Table 3: Measures of independent information when using the set of sentiment variables Positive, Negative, Trust compared to Positive, Negative, Skepticism. Data is from the target McDonalds.

Sentiment Name	Linear Regression		Lasso Regularization	
	R2	crossval R2	R2	crossval R2
Frequency	0.00738	-0.0275	-0.000	-0.0412
Buy	0.0729	-0.0408	0.079	-0.476
Sell	0.156	0.050	0.156	0.050
Financial Increase	0.00205	-0.287	0.0	-0.0291
Financial Down	0.185	0.0928	0.185	0.0980

Table 4: Explained variance R2, 50-50 crossvalidated R2 for linear regression, with and without Lasso regularization, for daily Bitcoin sentiment variables.

Metrics for the independent information in the sets of sentiment variables Positivity, Negativity, Trust and Positivity, Negativity, Skepticism compared for the target McDonalds can be seen in Table 2.

The bootstrapped confidence intervals for the differences of the metrics can be seen in Table 3.

4.5 Regression to price data

The correlation plot between the sentiment variables and the price change of Bitcoins can be seen in Figure 9.

The results of the presented regressions are presented in Table 4. The 13 sentiment variables that are not shown all had a negative explained variance for unseen data, i.e. their prediction was worse than no prediction at all.

5 Discussion

In this section the results of the analysis performed above will be discussed. Exploratory analysis that has not been described in detail will also be mentioned, and qualitative discussion and conclusions will be presented.

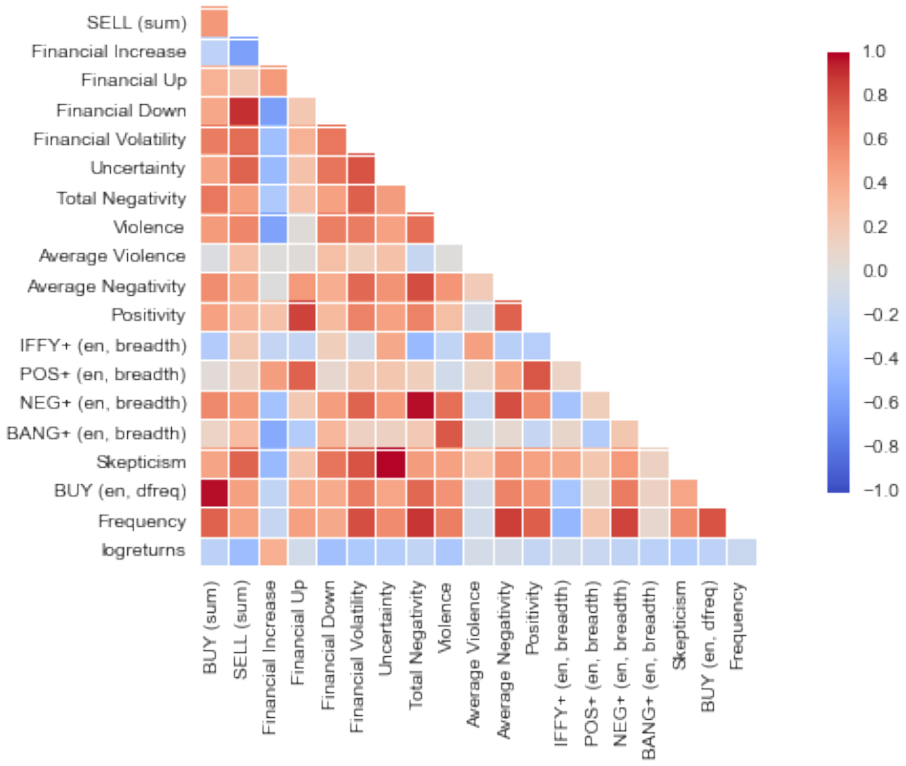


Figure 9: Bitcoin corrplot daily resolution

5.1 Limitations of the setup

The aim of this work has been to find a method that selects a subset of possible sentiment variables that contain as much independent information as possible. The main limitation with the problem formulation is the ambiguity of information in a random variable.

Implicit assumptions on how the sentiment variables are distributed have therefore been made made, and information for a set of random variables has been considered as some measure of unpredictability *between* the random variables.

Below a few limitations for measuring information with the setup and the data stream at hand will be listed. Understanding these limitations are important for future work and for working with the sentiment variables in practice.

Overall it is the case that analysis of measurements without considerations about the underlying process generating the sentiment, or even the distribution of the measurements, make inference from them suboptimal.

To estimate distributions or make theoretical assumptions about the underlying processes is difficult in this case, since the sentiment variables computed by Gavagai measure very many different types of documents that are all aggregated to one number. The documents types (i.e. if they are articles, blog posts or tweets, etc.) as well as the language for discussing them changes rapidly over time.

Below more specific points are listed regarding the data stream that allow for improvement:

1. The misclassification rate of the sentiment variables are unknown which leads to that the sentiment variables contain an unknown amount of possibly correlated noise.
2. The classification of sentiment in a document is done Naively, i.e. without modeling how combination of the features (keywords) predict a sentiment.
3. The classification does not model the probability for specific combinations of sentiments being expressed in the same documents.
4. The evidence for classifying a sentiment could be used more optimally by distinguishing the predictive strength of different keywords.
5. Long documents increase the probability of containing sentiment specific keywords, and hence the contribution from one document to a sentiment variable can depend more on the length and less on content. The length of documents varies over order of magnitudes over different documents.

A potential way to improve these shortcomings is proposed in section 5.7.

Since very little can be known about the sentiment variables without such improvements, the only way to draw conclusions from the data is to have relatively weak assumptions regarding their distributions.

The methods described above relies on the assumptions of stationarity and normality, which are tested with the Ljung box Q-test and the Augmented Dickey Fuller test.

5.2 Specific use cases for implementation

The outset of this work was to investigate how to select sentiment variables for different types of problems. The most important needs for the three use cases Political prediction, Financial prediction and Business intelligence were discussed in section 1.3.

Political prediction

When measuring sentiment with keywords for political sentiment there are vast amounts of noise from different types of sources, such as bloggers, news articles and discussion forums.

If a formal prediction of the political outcome should be performed, a more complex modeling of the underlying individuals that publish their political sentiment as digital documents should be used - counting individuals rather than documents or keywords - ought to be used as a proxy for the entire population of voters.

Exploratory work was performed to test if the sentiment variables at hand were able to predict the daily change of the Swedish voting polls. This was done by aggregating voting polls from Demoskop, Ipsos, Novus, SCB, Sifo, YouGov and United Minds with a multivariate local level model [19] to pool the results to a best estimate summing to 100% of voters, similar to the poll-of-polls as suggested in [8] (the implementation is attributed to Måns Magnusson [15]).

The daily relative change of this aggregated polls of voters was then used as endogenous variable in a linear regression with the sentiment variables Positive, Negative, Skepticism, Trust and Violence for each political party as the exogenous variables. The exploratory work did not support any clear pattern, and not much could be said mainly due to lack of data points. The topic could therefore not be investigated further.

Financial prediction

When using sentiment analysis for financial sentiment prediction an issue is that the documents that are being analyzed are in general weakly related to the investor or analysts expertise, and their market impact.

Although it has previously been showed that sentiment analysis can predict

financial markets, the results above can not verify this.

For future work there are possible improvements for verifying the hypothesis. The sentiment analysis could be segmented by the source of the documents to be treaded individually, the sentiment classification could be improved, the documents used could be weighted by their influence or expertise, or the regression model used could be changed.

Business intelligence

One Key Performance Indicator that could be possible to estimate is the impact or *engagement* of marketing campaigns. This by monitoring the short term changes and measuring both the magnitude and the sentiment of the change. For many business targets there is a large pool of documents that can be used to measure sentiment towards them. For publicly traded companies however, many documents regarding them are not from customers but investors and should be carefully divided.

Summary

For any practical use case, segmenting sentiment analysis results based on type of document and type of source is useful. This is helpful both for humans as well as for prediction when using sentiment variables. Prediction of Swedish political results did not yield any promising results task with the data stream at hand and without taking the documents sources into account.

5.3 Predicting the SP500 Volatility Index

From the results the assumptions of equal variance, stationarity, independence of residuals could not be rejected by the tests performed. Some deviation from the assumptions of residuals being normally distributed was observed.

The Granger causality test could not reject the hypothesis of non-causality. Furthermore the amount of explained variance was very low when benchmarking with the autoregressive model, especially when performing cross validation.

The results might be improved by using more data points, segmenting sentiment analysis by the source of the documents, improving sentiment classification, or weighting sentiment by the rank of expertise or influence before aggregating the sentiment.

For the regression model itself, future improvements could be to model the volatility as a latent variable in a state space model. This might remove any dependence between the residuals of the predicted volatility and time which could not otherwise be ruled out.

5.4 Anomalies

The anomaly detection method proved to be useful to filter out outlier-events that are not suitable to be used when fitting a regression model, or to just detect anomalies. It used relatively few parameters to estimate the decision boundaries.

Detecting anomalies in real time data was not done. This would require the parameters of the decision boundaries would have to be estimated for every new data point for a future time horizon T.

The method of constructing confidence intervals is equivalent to using the one sided Mahalanobis distance [16] assuming that the correlation between the variables are zero. Dropping the assumption of zero correlation is a possible generalization, but would require more parameters to be estimated.

Another alternative method that could be preferred if very many data points would be available is to use the empirical distributions quantile bounds for the confidence intervals, which drops the assumption of normality.

5.5 Analyzing similarity in a subset of sentiment variables

The results from the three methods for selecting sentiment variables that explain as much independent information as possible all imply that Positivity, Negativity and Trust contain more independent information.

The confidence intervals for the differences however all contain zero so the null hypothesis that Trust and Skepticism are equally good can not be ruled out on the $p = 0.05$ significance level.

Showing significance in the tests might be possible with more data points.

5.6 Regression to price data

The sentiment variable which explained most variance of the price change of bitcoin for unseen test-data was the Sell sentiment variable, the Financial Down explained some variance as well. The other 16 sentiment variables could not decrease the variance for unseen data. Hence, using exogenous data to select sentiment variables was not very useful in this case. Having more data points together with a more sophisticated regression could perhaps explain variance for more than the two variables above.

The correlation plot between the sentiment variables of Bitcoin as seen above is in practice useful to select sentiment variables manually. For example Buy (dfreq), Financial Increase and Average Negativity seem not to be very corre-

lated compared to other variables.

5.7 Road map for implementing large scale sentiment analysis

In this section a practical road map for efficient implementation of sentiment analysis for an industrial setting is proposed. Recommendations regarding the shortcomings of the present design for sentiment analysis are suggested.

With the present design there are a few important questions to address. Each choice has its pros and cons and depends on the goal of the problem that is needed to solve.

1. What kind of documents should be treated?
2. How should sentiment in very long texts be measured?
3. How should sentiment in very short texts be measured?
4. Should copies of the same text be accounted for?

These questions are preferably decided individually for every context to get the best results. In cases where this is impractical, a straightforward method is outlined as follows:

1. For each document and each sentiment, predict the probability that the document shows the sentiment towards the studied target.
2. Aggregate the estimated percentage of documents that show each of the sentiments towards the target. Compute the confidence bounds of the percentages with
3. Present the percentage of each sentiment - with uncertainties - as well as the exact number of mentions of the target.

This procedure will take into account more of the information contained in the texts, and avoid diluting the sentiment measure with very uncertain sentiment classifications. This is because the sentiment for each document is taken as a continuous value so the information of weak classifications will not be lost, and uncertain classifications will be taken into account. This mathematically coherent way of handling multiple documents also suits better for stringent analysis or regression.

Moreover, the last step of presenting the number of documents as well as the percentages and uncertainties of each sentiment is a visualization that is arguably easier to understand. A weakness is that the strength of sentiment in

each document will not necessarily be expressed in the probability of the sentiment — the procedure would require an extra variable of e.g. *Strong positivity* to achieve this.

It should be pointed out that even if the sentiment classification algorithm is not so good, the confidence bounds of the percentage of sentiment expressing documents will become narrow if there are many document (as a result of the Law of large numbers).

The prediction of each sentiment can in its simplest form be done by manually using known sentiment keywords and estimating the probability for the sentiment that each keyword contributes. One recommendation for this is to evaluate the procedure with previously sentiment labeled data. A less simple method is to use such a dataset to fit a model for predicting sentiment.

Regression could be done with a simple method such as naive-Bayes or logistic regression predicting probability with the co-occurrence of with keywords used as features. More sophisticated method with recent success are deep recurrent neural network models such as has been demonstrated in [21].

To train a model, i.e. to fit the regression, a qualitative sentiment labeled dataset of both longer articles and shorter texts with as many relevant sentiments as possible is needed in the language of interest. This can for example can be taken from public review data labeled with sentiment.

From the predicted probability in step 1 above, the percentage of documents that show sentiment is simply estimated by summing over each relevant documents predicted probability. The confidence bounds of the the percentage can be computed by bootstrapping. This is done by randomly letting each of the sentiment for each document be fixed to either true or false, drawn independently for the predicted probability of the sentiment being true. Repeating the procedure many times yields the estimated distribution for the number of documents expressing the sentiment, and confidence interval can be constructed by simply sorting the bootstrapped estimates and taking the empirical quantiles.

Confidence intervals can also be computed for any kind of comparison that can be interesting, such as comparing different time intervals, different targets, the relative content of sentiment, or sentiment for different sources or demographics. This would be implemented with the same bootstrapping method as described above.

The problem that has been the focus of this work - selecting which sentiment variables to use - will be a bit simpler with the above model since the confidence bounds of each sentiment variable is known and is a good criterion for selection. It also enhances the capabilities for selecting sentiment variables through inspection, which is often preferred.

This work has evaluated three black box models or selecting sentiment variables that contain as much information as possible. The one that is most strongly recommended is the PCA based technique described above. Simply inspecting

the correlation between each pair of variable can also be useful to select a set of variables manually.

It can be emphasized that with any model for aggregating sentiment from the Internet it is good to have the implementation consider the specifics of the problem and goal at hand and engineer a solution around this. Then gradually be improve such a model so that the data is always reliable with few misclassification and a clear way to understanding the output is important.

5.8 Conclusions

Selecting sentiment variables containing as much independent information as possible by regression to exogenous data did not provide conclusive results with the datasets used. Predicting financial markets with the keyword based sentiment analysis could not be shown to work significantly with the models evaluated, which agrees with the efficient market hypothesis.

When no exogenous data was available to select sentiment variables the three proposed black box methods agreed for the selection task, but could not be shown to be significant on the $p = 0.05$ level. More data points would be needed to show significance.

A road map to improve the sentiment variables for easier understanding of the output, and better utilization of the information in the analyzed documents was proposed. The outline of the road map is to estimate the probability of each document having a sentiment towards the target, and aggregate the distribution of all the documents with various sentiment toward the target. The model for the predicted probability can take more of the information regarding the document and how people express sentiment into account, and can be trained with labeled datasets of sentiment.

Emphasis in this work is also put on analyzing sentiment variables manually through visualization and use human knowledge before using a black box model.

Moreover, sentiment analysis for different contexts have been explored and discussed. Data and the needs of different use cases suggest that sentiment analysis should be segmented, with respect to the documents type, source, reach and influence, and if possible manually engineered to be able to capture what is of interest to an analyst. The best practice for implementing sentiment analysis varies significantly with the context and problem goal.

Apart from what is proposed in the road map, future work with more data points or sentiment variables are segmented by source and type could provide better predictive results. A state space model such as introduced in 1.4 for the measured sentiment could also prove useful for prediction or measuring underlying sentiment. Such work would need more exact specifications of what data points are made by.

References

[1] H. Bu and L. Pi. Does investor sentiment predict stock returns? the evidence from chinese stock market. *Journal of Systems Science and Complexity*, 27(1):130–143, 2014.

[2] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[3] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[4] B. Exchange. Bitstamp data source, 2015.

[5] D. Gillblad and A. Holst. Dependency derivation in industrial process data. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 599–602. IEEE, 2001.

[6] D. Gillblad and A. Holst. Dependency derivation in industrial process data. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 599–602. IEEE, 2001.

[7] D. N. . P. Gujarati and D. C. Causality in economics: The granger causality test. *New York: McGraw-Hill. pp. 652–658. ISBN*, pages 978–007–127625–2., 2009.

[8] S. Jackman. Pooling the polls over an election campaign. *Australian Journal of Political Science*, 40(4):499–517, 2005.

[9] M. Jafari-Mamaghani. Non-parametric analysis of granger causality using local measures of divergence. *Applied Mathematical Sciences*, 7(83):4107–4236, 2013.

[10] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[11] J. Kaminski and P. Gloor. Nowcasting the bitcoin market with twitter signals. *arXiv preprint arXiv:1406.7577*, 2014.

[12] R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145, 1995.

[13] H. Lang. *Topics on Applied Mathematical Statistics*. Royal Institute of Technology, 2013.

[14] A. C. Lazo and P. N. Rathie. On the entropy of continuous probability distributions. *Information Theory, IEEE Transactions on*, 24(1):120–122, 1978.

[15] M. Magnusson. Botten ada, 2015.

[16] P. C. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India 2*, 2(1):49–55, 1936.

[17] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[19] G. Petris, S. Petrone, and P. Campagnoli. *Dynamic linear models*. Springer, 2009.

[20] J. Shlens. A tutorial on principal component analysis. *CoRR*, abs/1404.1100, 2014.

[21] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[22] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[23] H. Theil. Economic forecast and policy, vol. xv of contributions to economic analysis, 1961.

[24] J. A. Thomas and J. A. Thomas. *Elements of information theory*. Wiley New York, 2006.

[25] R. Tsay. *Analysis of Financial Time Series. Hoboken, NJ:*. Wiley & Sons, Inc., 2005.

[26] P. C. Wong and R. D. Bergeron. 30 years of multidimensional multivariate visualization. 1997.

6 Appendix

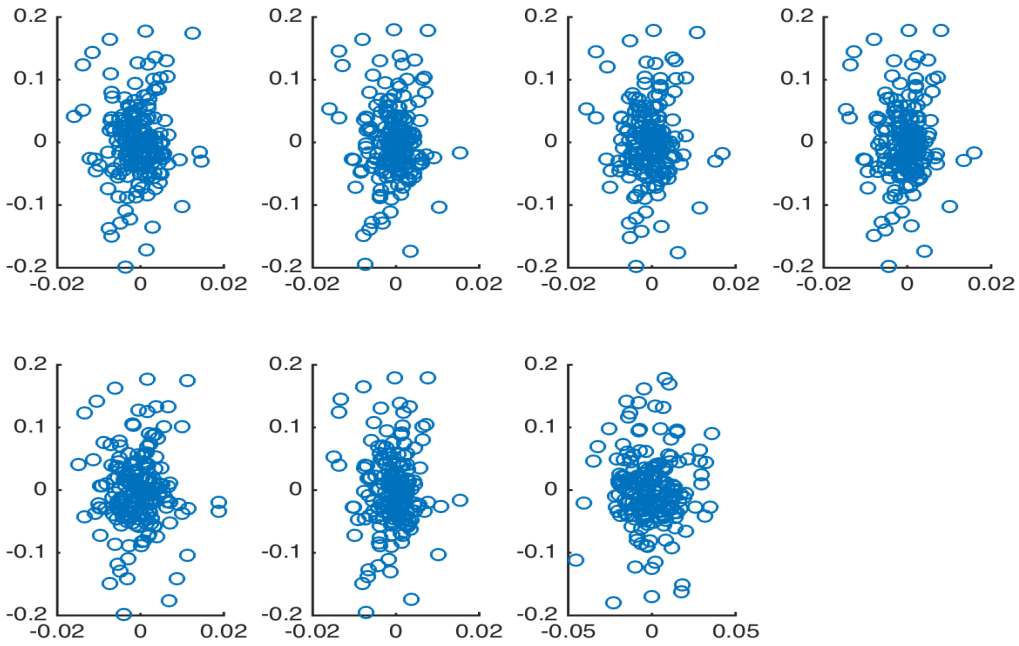


Figure 10: Residuals vs fitted values for the regression to the VIX. Does not let us reject assumption of equal variance.

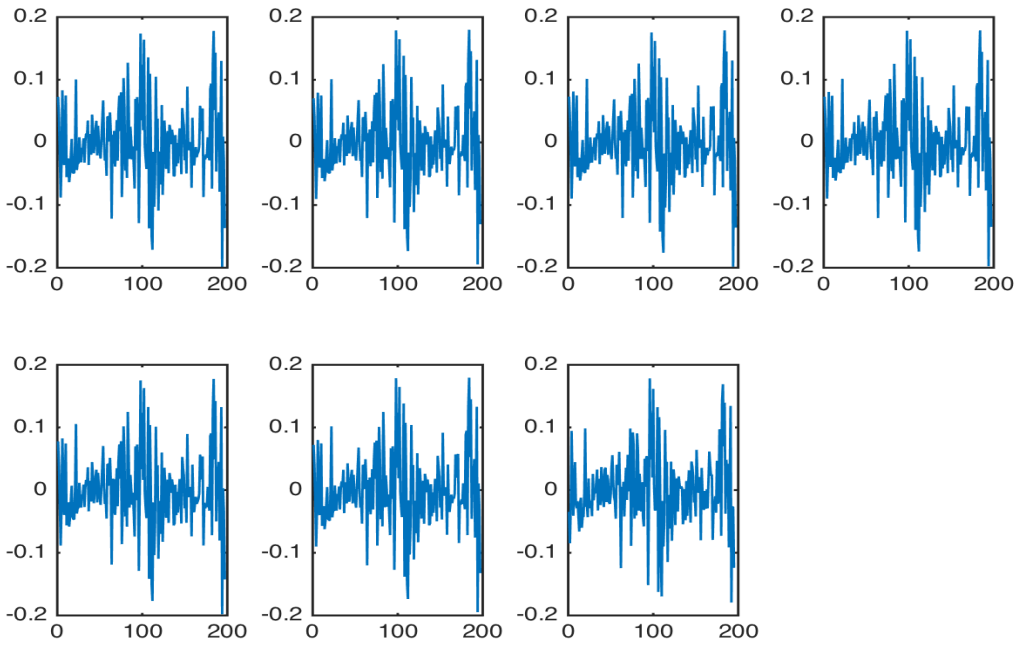


Figure 11: Residuals for each fit for regression to the VIX. Does not let us reject assumption of independence.

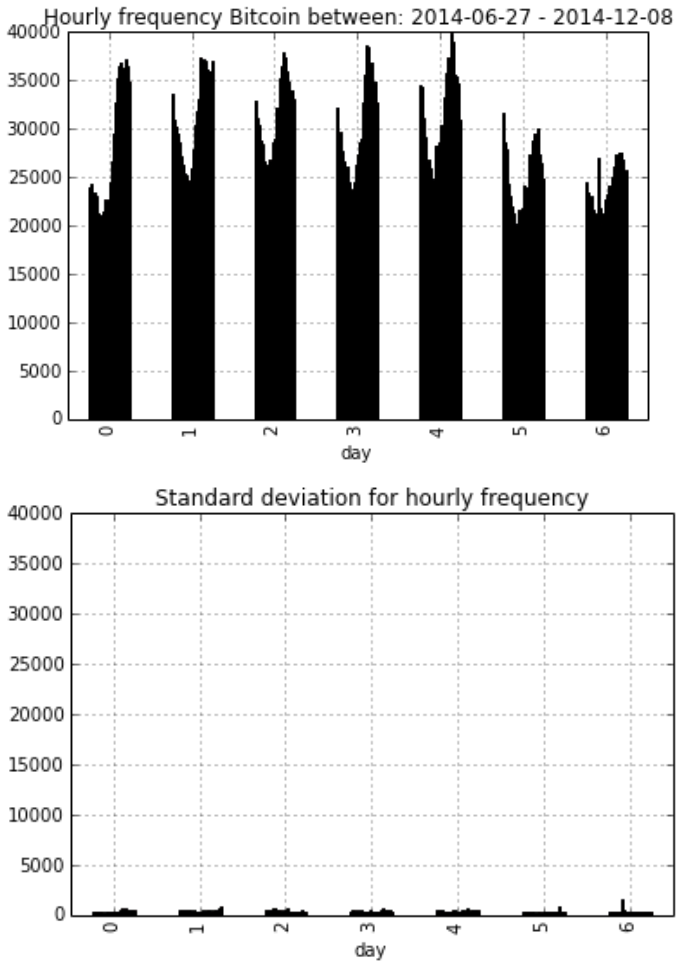


Figure 12: Hourly mention of Bitcoin Monday - Sunday, starting 00:00 CET.

	Negativity	Positivity	Trust	Skepticism
count	170.000000	170.000000	170.000000	170.000000
mean	15.182353	4.264706	0.682353	0.829412
std	52.754475	19.728796	5.438112	5.712310
min	-40.000000	-40.500000	-11.500000	-9.500000
25%	-12.875000	-8.500000	-3.000000	-2.875000
50%	0.000000	0.000000	0.000000	0.000000
75%	23.375000	15.625000	3.875000	3.500000
max	310.000000	82.000000	21.500000	23.500000

Figure 13: Descriptive statistics of the Mcdonalds dataset.

TRITA -MAT-E 2015:81
ISRN -KTH/MAT/E--15/81-SE