Anomaly detection on social media using ARIMA models

Tim Isbister
Abstract

Anomaly detection on social media using ARIMA models

Tim Isbister

This thesis explores whether it is possible to capture communication patterns from web-forums and detect anomalous user behaviour. Data from individuals on web-forums can be downloaded using web-crawlers, and tools as LIWC can make the data meaningful. If user data can be distinguished from white noise, statistical models such as ARIMA can be parametrized to identify the underlying structure and forecast data. It turned out that if enough data is captured, ARIMA models could suggest underlying patterns, therefore anomalous data can be identified. The anomalous data might suggest a change in the users’ behaviour.
Contents

1 Introduction 4
  1.1 Motivation ................................................. 4
  1.2 Background .................................................. 5
  1.3 Scope ........................................................ 7
  1.4 Outline ...................................................... 7

2 Related work 8
  2.1 Prediction with Social Media ................................. 8
  2.2 Just Google It! ................................................. 8

3 Fundamental concepts 9
  3.1 Definition ...................................................... 9
  3.2 Components ................................................... 9
  3.3 Functions ..................................................... 10
  3.4 Stationarity ................................................... 13
  3.5 ARIMA Model .................................................. 14
  3.6 Anomaly detection ............................................ 14

4 Dataset 16
  4.1 Semantic analysis ............................................ 16
  4.2 Creating feature vectors .................................... 16
  4.3 Creating a time series ...................................... 17

5 Box-Jenkins Methodology 18
  5.1 Stationarity ................................................... 18
  5.2 Identification ................................................ 19
  5.3 Estimation .................................................... 21
  5.4 Diagnostic checking ........................................ 22
  5.5 Forecasting ................................................... 23

6 Experiments 25
  6.1 Stationary analysis .......................................... 26
  6.2 Model identification ........................................ 26
  6.3 Model estimation ............................................. 28
  6.4 Model diagnostic checking ................................... 28
  6.5 Forecasts ..................................................... 29

7 Summary 34
  7.1 Discussion ................................................... 34
  7.2 Conclusions ............................................... 34
  7.3 Future work ................................................ 35

A
  Word list ....................................................... 36

B
  Dickey-fuller Table ............................................ 36
1 Introduction

The ability to freely express ideas and discuss a broad range of topics with like-minded people is a cornerstone in a democratic society. New forms of social media allow users to spread their views rapidly to a large group of people. Most people use the Internet and social media for harmless interactions, communication and for finding information but in some cases the Internet can serve as a forum for violent extremism and terrorism. It has been showed that Internet has played a prominent role in many terrorist attacks such as in the 2009 Fort Hood shootings, the 2008 Mumbai attacks, and the 2004 Madrid bombings, and in various terrorist plots, including amongst the Netherlands' Hofstad Group, and Colleen La Rose (Jihad Jane) and others plotting to murder the Swedish cartoonist, Lars Vilks [1].

Web forums are one place for extremist groups to spread their views and communicate. One example of a extremist web forum is Stormfront. Stormfront is the largest white supremacist Web forum in the world with close to 300,000 registered users. According to a report from the Southern Poverty Law Center (SPLC) [2], Stormfront members have murdered almost 100 people in the past six years. This is something that makes monitoring the Internet and social media an important task for intelligence analysts. The importance of the task is also identified by Europol that recently launched the European Union Internet Referral Unit (EU IRU) with the aim to combat terrorist propaganda and related violent extremist activities on the Internet.

By analysing web forums and keep track of users that might pose a threat towards the society there is a chance to prevent future attacks.

1.1 Motivation

The purpose of this thesis is to study whether there is an increased fixation of a radical topic on an Internet forum using time series analysis. This series can be thought of modeling how a person communicates on a social media platform, in the sense that the variable which is studied will describe the intensity of the users’ activity over a period of time.

When an adequate model is obtained for the series, the model can be viewed as a tool describing a possible pattern of the user, or more formally, the internal structure of the time series. The model will be able to produce forecasts for the intensity of the upcoming posts.

However, if the real data does not conform to the expected pattern suggested by the model, this will be viewed as anomalous and can indicate a change in the communication pattern, which might indicate a change in the users’ behaviour.

Single outliers will not be regarded as a behavioural change, rather unexpected bursts of activity, which might suggest that a person on the Internet is becoming more fixated on a particular topic.

Finally, the accuracy of the model’s forecasts will be examined, whether a generalized data independent model can be achieved, and whether these techniques are viable for real-life applications.

---

1 The chosen topic will be presented in Section 4
2 In data mining, anomaly detection is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset.
1.2 Background

Data obtained from observations collected over time is very common. In business it is possible to observe weekly interest rates, daily closing stock prices, monthly price indices and yearly sales figures. In biological sciences, electrical activity of the heart at millisecond intervals [3, p. 1].

Data where time is the independent variable is called a time series. Analysis of time series is what characterises this thesis and a specific model will be reviewed in detail and tested. In general the main purposes of time series analysis is twofold:

- to understand or model the stochastic mechanism that gives rise to an observed series
- predict future values based on the history of a series and, possibly, other related series or factors

Studying models that incorporate dependence is the key concept in time series analysis. The analysis accounts for the fact that data points taken over time may have an internal structure that should be accounted for [4].

Figure 1 shows a graphical example of a time series. It’s a measurement of the monthly air traffic volume from 1996 to 2014 in the US. The series has plenty of data that can be applied to manifest the seasonal variation, trends, autocorrelation\(^3\) and so forth, which can be used to construct a model that estimates the internal structure.

When such a model is created using historical data, it can then be used to generate future values for the series, known as forecasts.

The time series model that will be addressed in this thesis was developed by the statisticians George Box and Gwilym Jenkins, and is widely known as Box-Jenkins method [5] or the autoregressive integrated moving average model, and it has gained great popularity since the publication of *Time Series Analysis:*

---

\(^3\)The correlation with the same variable at a two different time points; such concepts will be described in detail in Section 3.
1. Make Stationary

2. Identify Model

3. Estimate Model Parameters

4. Use model for forecasting

Diagnostic Checking

Is the model adequate?

No

Yes

Figure 2: Box-Jenkins Approach

*Forecasting and Control, 1970*. Their methodology consists of the following four-stage iterative approach represented by Figure 2.
1.3 Scope
The time series that will be studied in this thesis will consist of the observations of one variable, a univariate time series, within a discrete time period.

The semantic analysis function which provides a score of a user’s intensity is a simple function, more advanced semantic analysis is a direction that should be considered for future work.

1.4 Outline
In Section 1, a brief introduction is made to the topic, and a motivation of our problem is stated. Section 2 review related work. Latter in Section 3, fundamental concepts about time series analysis are introduced.

The dataset will be presented in Section 4, and explained in some detail how its fitted to a time series. The flowchart seen in Figure 2 will be used as a blueprint for Section 5 and 6. Where Stage 1 corresponds to Section 5.1 and Section 6.1, Stage 2, Section 5.2 and 6.2 and so fourth. Section 5 is the theoretical part of the Box-Jenkins Methodology, and Section 6 is the actual implementation of it, where the results of the experiments are presented in Section 6.5.

Finally, the results from Section 6.5 are discussed in Section 7.1, and conclusions about the questions presented in Section 1.1 are given.
2 Related work

2.1 Prediction with Social Media

Using social media to predict the future has been done in many different ways. Some work has focused on prediction the outcomes of elections using social media. In [6] the focus is on the relation between social media and public opinion and in [7] several forms of social media is used in combination with ARIMA models to predict the outcome of the UK election 2010. [7] illustrates the predicting power of common people when their data is aggregated.

To quote, ”boundedly rational individual[s] are capable of making, all together, a near-to-optimal decision, often outperforming every individual’s intelligence, meaning that the crowd, taken as an intelligent entity, is smarter than most of any human counterparts taken singularly.” [7, p. 58]

This idea, of aggregating individual data to a new data set will be considered in our experiments in Section 6.5.

2.2 Just Google It!

Since the accuracy of forecasts is of great importance when trying to find anomalous data, a master thesis in Norway of the research in economics and management focused on improving forecast accuracy.

The study about forecasting the Norwegian unemployment figures with the help of web queries is seen in [8]. It used the same statistical model as in this thesis, to forecast data. The time series unemployment data is a collection of eight years. They picked the most parsimonious models according to the same information criteria that’s explained in Section 5.2.

A significant improvement of the statistical ARIMA models is demonstrated. An ordinary ARIMA model is expanded with an external variable to improve the accuracy, named ARIMAX. The external variable is collected from www.google.com/trends/ where they collect data with key words such as: ”profice, adecco, manpower, top temp, toptemp, jobzone”. Google trends has public information about the amount of times each key word searched daily.

Once they normalized and prepared the data, it turned out that the forecast accuracy was improved with 19.8% by the ARIMAX models, compared to the pure ARIMA models. This is a technique that could be used in our context to improve forecast accuracy.
3 Fundamental concepts

This section introduces the core concepts in the theory of time series models, such as; Univariate time series, time series components, functions, anomalies, stationarity and the model itself.

3.1 Definition

A univariate time series is defined as a set of observations measured over time, in the sense that the measurements are collected sequentially over equal increments, of one variable. Where \( Y \) is the variable measured over a time period \( t \), written mathematically:

\[
Y_t : t = 1, 2, ..., n \in \mathbb{N}
\] (1)

The series can either be discrete or continuous, this thesis will focus on a discrete series where each observation is measured at discrete point of time. The measurements of time is usually recorded at daily, weekly, monthly, quarterly or yearly separations.

Since a time series can evolve in many different ways, it can also be called a stochastic process or a random process. A stochastic process is a process where only the starting point may be known, and there are infinitely many directions in which the process may evolve [12]. It can be viewed as the opposite to a deterministic process, that can only evolve in one way, e.g. a sine curve.

3.2 Components

There are different types of patterns that can be viewed by looking at a time series. The patterns may help understanding the internal structure and increase the quality of forecasts. Any time series can contain some or all of the following components:

**Trend (T)**
The trend can be seen as a long term change in the data, that can either be positive or negative.

**Seasonal (S)**
A seasonal pattern exists when there is a correlation between monthly, quarterly or yearly data.

**Cyclic (C)**
Cyclical variation describes the medium term changes in the time series which is caused by circumstances, that repeats in cycles. The duration extends over longer periods of time, usually two or more years.

**Irregular (I)**
Irregular variations in the time series are caused by unpredictable influences, which repeats in a particular pattern.

Considering those four components, two different models of time series are generally used:

\[
Y_t = T_t \ast S_t \ast C_t \ast I_t
\] (2)

\[
Y_t = T_t + S_t + C_t + I_t
\] (3)
Where the multiplicative model is based on the assumption that the components are not independent of each other, and might influence each other. The additive model on the other hand assumes that the components are completely independent of each other.

3.3 Functions

A stochastic process $Y_t$ has several functions that will be used to decide the parameters of our model in Section 5. First, the mean function which is defined as the expected value $E$ at time $t$.

$$
\mu_t = E(Y_t), \ t \in \mathbb{N}
$$

Notice the different way of expressing $\mu$ compared to standard statistics. $\mu$ is now a function dependent on time. Therefore $\mu$ can be different at each time point $t$.

Further, the auto covariance function $\gamma_{t,s}$ is defined as

$$
\gamma_{t,s} = \text{Cov}(Y_t, Y_s)
$$

where $\text{Cov}(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)]$, which expresses the covariance of the same variable at lagged time points $t$ and $s$.

The autocorrelation function $\rho_{t,s}$ that will be referred as the ACF, is given by

$$
\rho_{t,s} = \frac{\text{Corr}(Y_t, Y_s)}{\sqrt{\text{Var}(Y_t)\text{Var}(Y_s)}}, \text{ where, again measuring the same variable over time, in this case, the correlation of } Y_t \text{ and } Y_s.
$$

Important properties from the above mentioned auto covariance and autocorrelation functions;

$$
\begin{align*}
\gamma_{t,t} &= \text{Var}(Y_t) \\
\rho_{t,t} &= 1 \\
\gamma_{t,s} &= \gamma_{s,t} \\
\rho_{t,s} &= \rho_{s,t} \\
|\gamma_{t,s}| &\leq \sqrt{\gamma_{t,t}\gamma_{s,s}} \\
|\rho_{t,s}| &\leq 1
\end{align*}
$$

The $\rho_{t,s}$ at 1 indicates strong positive correlation, and at $-1$, strong negative correlation, and a value close to zero indicates weak dependence. If it’s equal to zero, it is said to be uncorrelated [3, p. 12].

The Random Walk

Another concept in time series is the random walk which is in fact a special case of our model that will be introduced in Section 3.5, and a modern model for describing stock price series. However, this will be used as an example series to investigate the properties of the functions in Section 3.3, also it will be referred to in several examples.

The random walk is defined as the sequence $\epsilon_1, \epsilon_2, \ldots \epsilon_t$ of independent, identically distributed random variables with zero mean and a constant variance.

$$
\begin{align*}
Y_1 &= \epsilon_1 \\
Y_2 &= \epsilon_1 + \epsilon_2 \\
& \vdots \\
Y_t &= \epsilon_1 + \epsilon_2 + \epsilon_t
\end{align*}
$$

\[A \text{ lag is an earlier data point in the series}\]
Where the next time point, or, value of $Y_t$ is a result of the previous step plus a new random step, it can be rewritten as

$$Y_t = Y_{t-1} + \epsilon_t$$  \hspace{1cm} (9)

where Equation 9 denotes that the current position on the left hand side $Y_t$ is a result of the previous position $Y_{t-1}$ and a random step $\epsilon_t$.

By investigating the functions on the random walk, we notice that the mean function $\mu_t$ is zero, due to the expected value of the series $Y_t$ is $E(\epsilon_1 + \epsilon_2 + ... + \epsilon_t) = E(\epsilon_1) + E(\epsilon_2) + E(\epsilon_t) = 0 + 0 + ... + 0$. With an similar proof [3, p. 13] $Var(Y_t) = t\sigma^2$ and $\gamma_{t,s} = t\sigma^2$, where we note that the variance and covariance increases linearly with time.

To investigate the autocorrelation, Equation 6 can be rewritten as

$$\rho_{t,s} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}} = \sqrt{\frac{t}{s}} \quad 1 \leq t \leq s$$  \hspace{1cm} (10)

where the following numerical values can provide some intuition of how the correlation behaves over time.

$$\rho_{1,2} = \sqrt{\frac{2}{3}} = 0.707 \quad \rho_{8,9} = \sqrt{\frac{8}{9}} = 0.943$$

$$\rho_{24,25} = \sqrt{\frac{24}{25}} = 0.980 \quad \rho_{1,25} = \sqrt{\frac{1}{25}} = 0.200$$

By the above pairs of autocorrelation, time points are more strongly correlated as time goes by, compare $\rho_{1,2}$ with $\rho_{24,25}$, and time points who appear at distant time points are less correlated.

Figure 3 represents one simulated random walk which might not clarify that the expected mean is zero, though it has to be viewed as one instance of all possible random walks.

A larger set of random walks is seen in Figure 4 where it becomes more obvious that the expected mean value is close to zero, and that the variance does increase over time. The increase of variance over time is due to the possibilities of paths increases as time goes by, one can compare time = 10 with time = 100 and a great increase is shown.

This idea of a time series with an increasing variance indicates that the behaviour of the time series is changing over time, and would be difficult to forecast, which leads us to the important concept of stationarity.
Figure 3: Simulated random walk

Figure 4: 1000 Independent random walks
3.4 Stationarity

Stationarity is an essential concept in the analysis of time series. The basic idea of stationarity is that the behaviour of the process does not change over time [3, p. 16].

It is said to be strictly stationary if the joint probability \( P(X, Y) \) of a series remains the same over time. Let \( F \) denote the joint probability, then

\[
F(Y_t) = F(Y_{t+k})
\]

the joint probability \( k \) periods apart should remain the same. However, strict stationarity is seldom the case with real data, hence the concept of weak stationarity is introduced, and will be used from now on when we refer to stationary. Weak stationarity refers to a process that the mean \( \mu \), variance \( \sigma^2 \) and autocovariance function \( \gamma \) are independent of time.

\[
E(Y_t) = \mu \ \forall t
\]

\[
Var(Y_t) = \gamma_0 \ \forall t
\]

\[
Cov(Y_t, Y_k) = Cov(Y_{t+s}, Y_{k+s}) \ \forall t, k, s
\]

The stationary series would look flattened compared to the non-stationary, it would have no trend, no seasonality, and constant correlation \( k \) lags apart [11]. Figure 5 represents an example of one time series that is non-stationary, which in
many cases can be very hard to tell by looking at it. Techniques for validating stationary are reviewed in Section 5.1. The second series in Figure 5 is the differentiated version of the above series, which is stationary.

The most common stationary series is the white noise process where each time point is totally random and has zero auto correlation, in contrast to the random walk. It is even strictly stationary since the joint probability distribution does not change over time. The white noise process is often used to describe series who has no pattern, and is referred to in Section 5, Figure 7, when diagnostic checking the model.

3.5 ARIMA Model

The acronym ARIMA\((p, d, q)\) represents the autoregressive integrated moving average model.

Where the first parameter \(p\) of the model denotes the order of the auto regression, which is the regression of the same variable, at lagged values. A \(p\)-th order AR\((p)\) process \(Y_t\) is expressed as:

\[
Y_t = \mu + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + \epsilon_t
\]  

Where \(Y_t\) is the stochastic process, \(\mu\) is the mean value, \(\alpha\) is the weight of the correlation coefficients that is multiplied with the lagged values of \(Y_t\). Finally an error term \(\epsilon_t\) is added, which is a independent and identically distributed random variable from a normal distribution with constant mean and variance, \(N(\mu, \sigma^2)\). The purpose of \(\epsilon_t\) is to represent everything new in the series that is not considered by it’s past values [3, p. 66].

The bottom line is that the auto regressive part is a linear combination of the \(p\) past values of itself, with an additional error term \(\epsilon_t\), and the mean \(\mu\).

The second parameter in the ARIMA model \(d\) represents the amount of differencing that needs to be done, if necessary, to achieve stationarity. It is said to be the integrated part of the model. A first order of differencing on the model would be written:

\[
\Delta Y_t = Y_t - Y_{t-1}
\]  

The final parameter \(q\) corresponds to the moving average which is a linear combination of it’s \(q\) past error terms written mathematically as:

\[
Y_t = \mu + \epsilon_t + \beta_1 \epsilon_{t-1} + \beta_2 \epsilon_{t-2} + \ldots + \beta_q \epsilon_{t-q}
\]  

It is structured similar as the autoregressive part, except the regression is made on the error terms, rather than the actual observations of \(Y_t\).

To combine the autoregression, integration, and moving average, the possible differentiation is made first, then the AR and MA equations are combined as ARMA\((p,q)\):

\[
Y_t = \mu + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + \epsilon_t + \beta_1 \epsilon_{t-1} + \beta_2 \epsilon_{t-2} + \ldots + \beta_q \epsilon_{t-q}
\]

3.6 Anomaly detection

The purpose of this thesis is to find anomalies using ARIMA models. Anomalies distinguishes from outliers in the sense that outliers are legitimate data points which is far away from the mean. Unlike outliers, anomalies refers to illegitimate
data points that are generated by another process which the rest of the data was generated by [15]. One might visualize the concept of another process as a possible behavioural change by the forum user.

To identify anomalies we will consider the observations who do not conform to the expected pattern suggested by the ARIMA model. The ARIMA model will provide confidence intervals with a certainty of 95%, where a burst of anomalies outside the confidence interval might suggest a person is becoming more fixated.
4 Dataset

The dataset that will be used for the time series is collected from the Swedish forum [https://www.flashback.org/](https://www.flashback.org/). It consists of an 8 year collection of 20 different users’ activity on the sub forum [https://www.flashback.org/f226](https://www.flashback.org/f226) “Integration och invandring”. The data was collected with the Python package Beautiful Soup.

4.1 Semantic analysis

In this thesis a linguistic analysis method is used, that is inspired by Linguistic Inquiry Word Count [13], which calculates the degree to which people use different categories of words across a wide array of texts.

Only one category is used in our case, which is defined as a word list created by Katie Cohen [6]. The used word list A consists of racist invective words, with references to an out-group, which is a group that the person does not identify himself with [14]. The forum users can be seen as members of an in-group [14] represented by the forum (in sociology an in-group is a social group that a person psychologically identifies as being a member of) and the people they discuss can be seen as an out-group.

For each of the 20 users in the dataset we have made a semantic analysis on the posts they have published in the forum.

At the beginning, the semantic analysis function normalized the posts, which turned out to give poor results, since short sentences with a majority of invective words would outscore a long post with a fair amount of invective words.

4.2 Creating feature vectors

For each user a feature vector $F_v$ is created. The feature vector can be described as:

$$F_v = [Date, F_1, F_2, \ldots, F_n]$$

Each feature $F_i$ in the vector represents a score provided by counting the number of words that occur in a users post who matches the words in the word list A. Each row in $F_v$ represents a post corresponding to a user which is stored in its unique .tsv-file. For example, 3 posts of a user is structured as following:

$$\begin{bmatrix}
\text{Date} & F1 \\
2014 - 07 - 06 & 0 \\
2014 - 07 - 10 & 2 \\
2014 - 07 - 15 & 4 \\
\end{bmatrix}$$

(19)

Where we note that there is only one feature that is of interest, since we are limited to a univariate time series, and that there are gaps between the measurements.

---

[5]Beautiful Soup is a Python package for parsing HTML and XML documents which is useful for web scraping

4.3 Creating a time series

To make the dataset work properly with the time series it is necessary that it is univariate and made up of measurements on equal time intervals [10]. The raw data that was downloaded with Beautiful Soup contained gaps between the measurements as seen in (19). Those gaps is due to a user could have been inactive on the forum.

A solution for the gaps is to provide zero-values for all days that no activity is recorded. This can be solved with SQL by joining our dataset with an dataset that contains zeros as the values for each feature $F_i$. Finally, to get monthly scores, all daily values where aggregated using R.

The final data that is used for the time series analysis would look like following:

$$
\begin{bmatrix}
\text{Date} & F1 \\
2014 - 07 & 8 \\
2014 - 08 & 14 \\
2014 - 09 & 18 \\
\end{bmatrix}
\tag{20}
$$

Where one aggregated value represents the monthly activity for the user.
5 Box-Jenkins Methodology

In this section the theory of the Box-Jenkins approach will be explained for building the model which emphasizes the principle of parsimony\(^7\) in choosing ARIMA models.

5.1 Stationarity

Dickey-Fuller test

In some cases it can be difficult to decide whether a time series is stationary or non-stationary by looking at it, hence the Dickey-Fuller test was developed [16] in 1979 by the statisticians David Dickey and Wayne Fuller.

It is used to quantify the evidence of non-stationarity, where the null hypothesis is that \( \alpha = 1 \) in a simple AR(1) process \( Y_t = \alpha Y_{t-1} + \epsilon_t \). Where \( \alpha \) is as before the correlation coefficient.

In other words it tests whether the stochastic process is a random walk as mentioned in Section 3.3. The test is structured as, it checks if \( \alpha \) lies on the unit-circle. If it does, its clearly non-stationary, since the correlation is to the lagged value in the process, and as seen in Figure 3, we know random walks are not stationary. The three possible cases when testing for a unit root are the following:

\[
Y_t = \alpha Y_{t-1} + \epsilon_t 
\]  

(21)

\[
Y_t = \mu + \alpha Y_{t-1} + \epsilon_t 
\]  

(22)

\[
Y_t = \mu + \phi_t + \alpha Y_{t-1} + \epsilon_t 
\]  

(23)

where \( \mu \) is the intercept, and \( \phi_t \) is a deterministic trend. However, in our data there are no trends, so the first two cases will be of our interest. In a simpler manner the hypothesis testing is expressed as:

\[
H_0 : \text{Non-stationarity } \alpha = 1 \\
H_A : \text{Stationarity } \alpha < 1 
\]

where the purpose is to reject the null hypothesis, which is achieved when the \( p \)-value\(^8\) is \( \leq 0.05 \), in other words, we reject the null hypothesis if we are outside the 95\% area of the normal distribution, which corresponds to 1.96\( \sigma \) [17].

Considering the simple AR(1) (random walk) (22) and (23) (random walk with drift), both of them will carried out the same way, since the test does not consider of \( \mu \) is zero or not. One can not simply test if \( \alpha = 1 \) we need to investigate if both \( Y_t \) and \( Y_{t-1} \) are non-stationary. The equation can be rewritten more neatly as:

\[
\Delta Y_t = \mu + \delta Y_{t-1} + \epsilon_t 
\]  

(24)

Where \( \delta = \alpha - 1 \), and \( \Delta Y_t \) is the first difference. Where we note that if \( \alpha = 1 \) the right hand side term vanishes. Implying the left hand side is stationary, since

\(^7\)A parsimonious model is a model that accomplishes a desired level of explanation or prediction with as few predictor variables as possible

\(^8\)The \( p \)-value or calculated probability is the estimated probability of rejecting the null hypothesis
it’s less than 1. On the other hand, the alternative hypothesis is true if $\alpha < 1$ it is within the unit-circle. To test actually whether we have a unit-root or not we can compare if the t-statistic from $\delta$ is less then values from the Dickey-Fuller distribution [21]. If it is, then the null hypothesis is rejected with 95% certainty.

**Augmented Dickey-Fuller test**

The problem with the Dickey-Fuller test is that it does not consider the autocorrelation in the error terms $\epsilon_t$, and is only valid for the $AR(p)$ process. With the augmented test, it generalizes the $AR(p)$ process to an $ARIMA(p,d,q)$ process [3, p. 79], which makes it possible to consider the amount of lags in the error terms.

The ADF provides a negative number, where the more negative the number is, the further out we are on the left skewed dickey-fuller distribution B, and rejecting the null with higher certainty.

To clarify an example is provided with the random walk series, which we assume will give a $p$-value greater than the significance level, and the ADF number should not be far from zero.

```r
Ran=rnorm(1000)
X=cumsum(Ran)  # Simulated a random walk
X=ts(X)
adf.test(X)  # Augmented Dickey-Fuller test
```

R provides the following data:

```
Augmented Dickey-Fuller Test
data: X
Dickey-Fuller = -1.9388, Lag order = 7, p-value = 0.6042
alternative hypothesis: stationary
```

where as expected the $p$-value is much greater than the significance level, and the ADF number is $-1.9$. Since this suggests that the series is non-stationary, it is necessary to differentiate it and re-investigate the test.

```
Augmented Dickey-Fuller Test
data: X
Dickey-Fuller = -8.0543, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

By differentiating the series once, we can now reject the null hypothesis, since $p$ is $\leq 0.05$. This would suggest an $ARIMA(p,1,q)$ model, and in general it is not common to differentiate more than twice in time series.

### 5.2 Identification

**Graphical identification**

Now that we have a method of deciding the integrated part of the $ARIMA$
model, techniques to build the autoregressive and moving average part will be reviewed.

The autocorrelation function, often referred to as the **ACF**, which was introduced in Section 3.3 will be used together with the closely related partial autocorrelation function **PACF** to decide the order of the AR and MA part. The functions will measure how observations in the series are related to each other over time. To measure the correlation it is necessary to plot the ACF and PACF against lagged values of the series.

As Equation 6 denotes, the ACF shows the autocorrelation between \( Y_t \) and \( Y_s \) for different values of \( s \). To simplify \( Y_t \) and \( Y_s \), can be written as \( Y_t \) and \( Y_{t-k} \), since the series is stationary, where \( k \) is the amount of lag.

So if \( Y_t \) and \( Y_{t-1} \) are correlated, then \( Y_{t-1} \) and \( Y_{t-2} \) are correlated. However, \( Y_t \) and \( Y_{t-2} \) will be correlated as well, since their both being correlated with \( Y_{t-1} \), or that they are directly correlated. To overcome this problem, the partial autocorrelation function is introduced, where all the points within the lag are partialed out. To decide whether to build a AR, MA or ARMA model, we follow these rules [3, p. 116]:

- **AR(p)** - the PACF drops off after lag p, and the ACF has a geometric decay.
- **MA(q)** - the ACF drops off after lag q, and the PACF has a geometric decay.
- **ARMA(p, q)** - geometric decay on both ACF and PACF.

To provide a graphical example, the following correlograms were obtained from the random walk. In Figure 6 we note that there is 1 significant spike in the PACF and the ACF has a geometric decay. Where this would suggest a
ARMA(1,0). Since within the PACF one value is larger than the dashed line, which should be considered. The dashed line is set as \( \frac{\sigma}{\sqrt{n}} \) for 95% confidence intervals, \( \sigma = 1.96 \) and \( n \) is the sample size.

On the other hand, it is not always possible to make a graphical identification of the model, since the ACF and PACF might not follow the mentioned rules. Hence, other specification methods will be reviewed as well.

**Regularization**

A more formal way of deciding the parameters to the model is to use the Akaike information criterion [3, p. 130]. The basic notion of Akaike information criterion is that by adding more parameters to the model it will fit better, but at the risk of overfitting, in the sense that information is lost about the real underlying pattern.

It will estimate the quality of a set of models relative to each other, in the sense that most parsimonious model will achieve the lowest score.

\[
AIC = \log(\frac{\sum \epsilon_i^2}{n}) + \frac{2k}{n} \tag{25}
\]

Where, \( \sum \epsilon_i^2 \) is the squared sum of the residuals, and \( n \) is the number of observations. \( k \) acts as the penalty parameter, it is the number of parameters fitted in the model. The purpose is to minimize the AIC, by decreasing the number of parameters \( k \) for e.g, the AIC will have a lower score, and the better fit.

### 5.3 Estimation

Now that methods for deciding the order of \( p, d, q \) in the ARIMA\((p,d,q)\) are reviewed, the last part is to estimate the parameters \( \alpha \) and \( \beta \) for \( p \) and \( q \).

There are several techniques to estimate the parameters where the most common ones are least squares and maximum likelihood estimation [3, p. 154], where the latter is used in R by default.

**Least Squares**

The purpose of least squares is to fit a line that minimizes the overall vertical distance of the residuals\(^9\) to the fitted line. It is defined as [18]:

\[
S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{26}
\]

Where \( y \) is the observed data, and \( \hat{y} \) is the predicted data. The smaller the sum \( S \) is, the better the line fits.

**Maximum Likelihood Estimation**

To the observed series, \( Y_1, Y_2, ..., Y_n \) the likelihood function \( L \) is said to be the joint probability density of gathering the data that was observed.

For \( ARIMA \) models \( L \) will be a function of the unknown parameters \( \alpha \) and \( \beta \). The maximum likelihood estimators are then defined as the values of the parameters for which the data actually observed are most likely to be. The actual math that calculates the MLE is advanced see [3, p. 158], and is automated by R.

\(^9\)A residual is the distance between the measured data, and the predicted data, \( \epsilon_t = y_t - \hat{y}_t \)
5.4 Diagnostic checking

At this point, methods of specifying the model and estimating the parameters are reviewed. It is time to test the goodness of fit for the model. If the suggested model fits poorly, modifications are suggested, see Figure 2.

There are two main approaches for model diagnostics [3, p. 175]. Either to analyse the residuals from the fitted model, or to check if the model is over-parametrized. The first approach will be used here because of simplicity.

A model is said to be well fitted if the residuals are uncorrelated. If there is a significant correlation between the residuals, then there is information left, which the model does not capture. Ideally, the residuals should behave like a white noise process, with zero mean.

Quantile-Quantile Plot

The first most simply step is to make a QQ-plot. Such a plot will display the quantiles of the data on the Y-axis, and the theoretical quantiles of a normal distribution on the X-axis. If the residuals are normally distributed, the QQ-plot looks approximately like a straight line. See Figure 7, which is what the expected residuals should look like in Section 6.4.

![Normal Q-Q Plot](image)

Figure 7: QQ-Plot from a simulated White Noise process
The Sample Autocorrelation Function
The second step to examine normality in the residuals is the sample autocorrelation function SACF, which is a slightly modified version of the ACF. The SACF is modified in the sense that it assumes stationarity, hence a common mean and variance is put into Equation 6 [3, p. 46].

As the previous ACF plots in Section 5.2, the purpose is to visually identify patterns, and preferably no patterns should occur within the residuals. However, as this can be a subjective approach, a more objective test is suggested as well.

The Ljung-Box Test
The more formal way of testing the residuals is the Ljung-Box Test. In addition to check the correlation between individual lags, this test checks overall correlation between a group of residuals. It may be that most of the residuals are uncorrelated, but some are close to their critical values. Brought together as a group, the ACF can seem excessive. The hypothesis are the following:

\[ H_0 : \text{The residuals are independently distributed} \]
\[ H_A : \text{The residuals are correlated with each other} \]

Where the test is defined as:

\[ Q = n(n+2) \sum_{k=1}^{h} \frac{\hat{\rho}_k^2}{n-k} \]  

(27)

\( n \) is the sample size, \( \hat{\rho}_k \) is the sampled autocorrelation at lag \( k \), and \( h \) is the amount of lags. For the \( H_0 \) to be valid, the test \( Q \) is supposed to follow a chi-square distribution with \( df \) degrees of freedom. With the ARMA(p,q) model in mind, the degree of freedom is set as \( h-p-q \).

5.5 Forecasting
The final, and probably the most important stage of the Box Jenkins methodology is forecasting. There are two broad types of forecasts, where one step ahead forecasts are generated for the next observation only and multi-step ahead forecasts are generated for 1, 2,.., \( s \) steps ahead.

When building models it’s common to use a portion of the data as a training data, and the latter 20% as test data. Then the test data can be used to measure how well the model forecasts [19].

To measure how well a model forecasts, it’s invalid to look how well it performs on historical data. Genuine measuring of forecasts should only consider the test data (hold-out set).

If our dataset is denoted as \( Y_1, Y_2, \ldots, Y_T \) the training data is set as \( Y_1, \ldots, Y_N \) and the test data as \( Y_{N+1}, \ldots, Y_T \). A \( h \) step ahead forecast can be written as \( \hat{Y}_{N+h|N} \). Which is the forecast \( N \) to \( h \) given the training set \( N \).

The forecast errors \( \epsilon_t \) is the difference between the actual values, and the forecasts provided with only the training data. According to [19] the Mean absolute error and Root mean squared error should be used to measure the accuracy if the forecasts are on the same scale as the prior data.
**Measuring forecast accuracy**

The mean absolute error is defined as: \( mean(|e_i|) \), and the Root mean squared error is defined as: \( \sqrt{mean(e_i^2)} \), where the smaller the mean or squared value of the forecasting errors, the better forecast. The following points are of importance:

- A model which fits the data well does not necessarily forecast well.
- A perfect fit can always be obtained by using a model with enough parameters.
- Over-fitting a model to data is as bad as failing to identify the systematic pattern in the data.
6 Experiments

This section will perform the Box-Jenkins Methodology on our dataset, where the reader might skip directly to Section 6.5 to see the actual results. A few of the twenty users’ will be picked randomly, and the Box-Jenkins Methodology will be applied, and forecasts about the users intensivity is presented.

User data

To begin with, one of the twenty users is picked at random\textsuperscript{10} and the corresponding time series is seen in Figure 8. The data seen in Figure 8 is used as training data for the model. It is a collection of monthly values from January 2008 to December 2013.

Where we note that the used word list did result in a higher activity the first year. The high values in 2008 should not be viewed as outliers, because as mentioned in Section 4, each data point is a aggregated value of a monthly collection. To clarify, the maximum value 59 found in July 2008, is the summation of 2, 1, 0, 3, 1, 1, 5, 3, 10, 14, 3, 0, 2, 0, 0, 1, 6, 0, 0, 0, 1, 0, 1, 2, 1, 0, 1, 0, 1, this might suggest that the word list had a lot of matches, or that the user actually was more active in 2008. The latter data from January 2009 to December 2013 is more stable, with a mean value of 5.9, compared to 26.6 within 2008.

\textsuperscript{10}Not purely random, since a few of the users had meaningless data
6.1 Stationary analysis

By viewing the time series in Figure 8, we note that there seems to be no upward or downward trend in the data. Which is one of the arguments that supports the data is stationary. However, be measuring the mean and variance manually over a fair amount of lags in the latter data, its unclear if its constant or not. The Augmented Dickey-Fuller test is therefore carried out.

\begin{verbatim}
Augmented Dickey-Fuller Test
data: User1
Dickey-Fuller = -2.5321, Lag order = 4, p-value = 0.3576
alternative hypothesis: stationary
\end{verbatim}

Where the above \( p \)-value suggests that we cannot reject the null hypothesis, which suggests that the series is non-stationary. By differencing once, the following output is given:

\begin{verbatim}
Augmented Dickey-Fuller Test
data: User1
Dickey-Fuller = -5.3058, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
\end{verbatim}

Where this suggests that the series is stationary, and that we can reject the null hypothesis with 95% certainty, and conclude that and \( ARIMA(p, 1, q) \) model is required.

6.2 Model identification

Graphical identification

It is necessary to plot the ACF and the PACF from the differentiated data to determine the \( ARMA \) parameters. By examining Figure 9 and Figure 10 it hints of the first pattern mentioned in Section 5.2. Where this would suggest an \( ARMA(2, 0) \) model, with the differentiated data. Though, this has to be viewed as a tentative model, which needs to be confirmed using the information criteria.

Information criteria

Once a set of \( ARMA \) models have been suggested, the information criteria can be used to provide more insight to the performance of the models. A matrix with a set of models is seen in Table 28, where we note that our tentative model on row zero, column two, has the lowest AIC of them all.

\[
\begin{bmatrix}
ARMA(p, q) & 0 & 1 & 2 & 3 \\
0 & 601.96 & 565.36 & 564.22 & 566.0 \\
1 & 569.89 & 564.8 & 566.35 & 568.0 \\
2 & 565.45 & 566.35 & 567.43 & 568.55 \\
3 & 565.7 & 567.02 & 568.36 & 569.84 \\
\end{bmatrix}
\]

Where this supports the suggested model \( ARMA(2, 0) \) from the graphical identification.
Automatic model selection

The same produce is used considering seasonality and trends, where a seasonal ARIMA model contains AR and MA data at times that are multiples of the span of the seasonality $s$.

$$ARIMA(p, d, q)_{s(P, D, Q)}$$  \hspace{1cm} (29)

Where $P$, $D$, $Q$, are the ARIMA parameters repeated over the seasonality $s$. A constant drift, can be added, if original series would indicate a trend.
Fortunately, this procedure of checking all nearby ARIMA models, including the ones with seasonality, and trends can be automated by R using the forecast package [22].

The `auto.arima` function returns the best fitted model according to the information criteria, and an ARIMA(2, 1, 0)x(1, 0, 0)_{12} was suggested, with an AIC = 482.34, that fits better according to the AIC compared to the previous model. The model selected by the automatic approach suggests that the auto regressive parameter seems to be correlated yearly, and will be discussed if it’s reasonable in Section 7.

### 6.3 Model estimation

This is an automated procedure by R using maximum likelihood estimation. However, the starting values for the parameters can be set as the values obtained from the ACF and PACF plots, but are recalculated for more accuracy with maximum likelihood estimation.

```
arima(x = User1, order = c(2, 1, 0))
Coefficients:
ar1     ar2
 0.6198  0.5092
s.e.    0.0916  0.0922

sigma^2 estimated as 44.75: log likelihood = -279, aic = 564.22
```

This yields the following differentiated ARMA(2, 0) as: \( Y_t = 0.6198 Y_{t-1} + 0.5092 Y_{t-1} + \epsilon_t \).

### 6.4 Model diagnostic checking

To investigate if the manually suggested ARIMA(2, 1, 0) succeeds fitting the time series from User1, there should be no significant correlation between the residuals. The most straightforward way of deciding if the residuals looks like white noise is to view the QQ-plot. In figure 11 it seems that the residuals are close to a normal distribution, since the majority of them fits the straight line. There are a few outliers probably because the data where a lot higher in the first year. Where the most significant errors of the fitted model are found. If the same model is fitted to all the years after the high activity period, we note in Figure 12 that the residuals are more normally distributed.

The Ljung-Box test is also carried out of the manually suggested model. We observe a very high p-value, which indicates we cannot reject the null hypothesis, so the residuals seems to be independent of each other.

```
Box-Ljung test
data:  resid(User1)
X-squared = 12.39, df = 22, p-value = 0.9285
```

\(^{11}\)Almost all nearby models, since in the current version of auto.arima, the differentiated models with drift are not considered unfortunately.
6.5 Forecasts

Forecasts from *User1* is presented with the manually picked $ARIMA(2, 1, 0)$, and the automated $ARIMA(2, 1, 0)x(1, 0, 0)_{12}$ model by *auto.arima* from Section 6. Both models were trained on the data from Jan 2008 to Dec 2013, and a forecast is presented in Figure 13 and Figure 14.

Another, seasonal model is also presented in Figure 15 from *User2*, to demonstrate the possible efficiency of seasonal models.

Finally, the idea of anomaly detection is presented in Figure 17.
6 Months forecast from ARIMA(2,1,0)

Figure 13: Non seasonal model from User1 compared with real data

<table>
<thead>
<tr>
<th>User1 ARIMA(2,1,0)</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>7.198065</td>
<td>4.6923880</td>
</tr>
<tr>
<td>Test set</td>
<td>1.260333</td>
<td>0.9402743</td>
</tr>
</tbody>
</table>

A close up in Figure 13, where the real data in the dashed line stayed within the confidence intervals. The RMSE from the six month test set is $\approx 1.26$. 
6 Months forecast from \textit{ARIMA}(2,1,0)(1,0,0)[12]

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure14.png}
\caption{Seasonal model from User1 compared with real data}
\end{figure}

<table>
<thead>
<tr>
<th>Users</th>
<th>ARIMA(2,1,0)(1,0,0)[12]</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training set</td>
<td>6.930807</td>
<td>4.559618</td>
</tr>
<tr>
<td></td>
<td>Test set</td>
<td>1.781860</td>
<td>1.375320</td>
</tr>
</tbody>
</table>

Here we note that the automated seasonal model performed worse with the test data compared to the manually picked model, however it performed better with the training data, which is the reason for the lower AIC. The seasonal model in
Figure 15: Seasonal forecast User2

Figure 15 seems to capture the overall movement of the time series. User2 was one of the most active users, with a significant amount of data compared to the others. Figure 16 shows the overall movement of the 20 users values aggregated to one, which could be seen as the movement of the crowd, as suggested in Section 2.

Anomaly detection
In Figure 17 a forecast which found anomalies is presented. We note the real data is outside the confidence intervals for the two first months and should might be considered as anomalous data.
Figure 16: Non seasonal model for the crowd

Figure 17: Automated model with anomaly data
7 Summary

7.1 Discussion

The ARIMA models seem to capture the data within the confidence intervals in the majority of the cases when making short term forecasts from several users. On the other hand, it is preferably to do only a one-step ahead forecast, to be more confident of accuracy for real applications. Since in most cases there seems to be no seasonal pattern, and the non seasonal forecasts converges to the mean value in the long run.

Automated models picked by the lowest AIC would not necessary produce more accurate forecasts as seen in Figure 14. Therefore the automated models with the lowest AIC might not be the first choice, an extension for automating the most accurate models with the test data should be considered.

But is it really true that there is a real seasonal pattern within the data, or might it just appear to be seasonal in the way we generate the data, or that it happens to fit the training data. It might not be real seasonal data, since when the same user data was generated with normalization it did not turn out to be seasonal. According to [20] seasonal data is most common in economic time series, which is not the case.

If no pattern can be seen in the data, which turned out to be true in many cases since its close to weak white noise, forecasts can however still be made, and anomalous data can be detected.

7.2 Conclusions

Data dependent model

The statement in Section 1 of whether it is possible or not to make a generalized ARIMA model for all users’ turned out to be false, since the time series differ to much to generalize one ARIMA model with the same parameters. However, it seems like if it is possible to fit a seasonal model, it can capture the overall movement of the series better than the non-seasonal, and produce better forecasts in the long run.

Behavioral change

As seen in Figure 17 the real data is out of the predicted scope from the model. Does this data argue that the user is becoming more fixated on a particular topic, or that the word list just provided higher scores.

Since a high score is related to the amount of words, it could mean many things, such as:

- The user is more active
- The user is triggered by other users’
- World events influence forum activity
- Seasonal effects, such as climate

Or finally,
- User is more radical than considered normal.
By manually checking posts that provided high scores with the non-normalized data, one can not reject that the posts are long, and that there is a lot of thought behind them. But, should it be considered as abnormal behaviour or, that the user just has more time to express himself.

This research can not draw any conclusions about if the users’ behaviour did change or not, since this is a another field of science.

7.3 Future work

Automated anomaly detection
All the anomalous data that was found in this thesis was in a predefined data set. In reality the detection of anomalies would preferable work with new incoming data. An idea to build further on would be to download new data in real time with a web-crawler in Python for e.g. The reason for using python with this particular task is that it has support for ARIMA models [23]. An automatic ARIMA model could then be built, using AIC for example and the highest values of the forecasted data would be compared to real incoming data. If the real values are above the forecasts for a certain period of time, this could be indicated to the utilizers of the program. This software could then be parallelized to several cores or computers, to simultaneously detect anomalies on several users’.

Improved semantical analysis
We have to admit that the current semantical analysis is rather simple. It does not consider the way of how people write, and it would be naive to indicate a behavioural change by a high amount of words that matches our word list.

For e.g a person who is not considered as radical might score high with our function because he is using the invective words in another context. How likely it is to find such users’ on the targeted forum is a subjective issue, that has to be future analysed.
A

Word list

Afrikan  Afrikaner  Afrikanerna  Arab
Araber   Araberna  Babbe   Blatte
Blattn   Babbe    Blatte   Blatte
Bidragsturist   Bidragsturister   Bidragsturisterna
Bluneger  Blanege   Blanege   Blanege
Flyktingbarn  Främmende  Främling  Främlingar
Kamelknullare  Kamelknullarna  Kamelknullare  Kamelknullare
Kulturberikare  Kulturberikarna  Kulturberikarna  Kulturberikarna
Kurdarna  MENA    Menafolk  Menafolk
Menapack  Muslin  Muslimer  Muslimerna
Musselmanen  Musselmaner  Musselmanerna  Musselmanern
Musselmansk  Mörkhyad  Mörkhyade  Neger
Negern    Negern    Negern    Negern
Negroida  Nysvensk  Nysvenska  Nysvenskar
Nysvenskarna  Rasfrämling  Rasfrämlingar  Rasfrämlingarna
Rom      Romzie  Rommer  Romers
Sandneger  Sandnegrerna  Skäggapa  Skäggapor
Skäggapor  Skäggapor  Skäggapor  Skäggapor
Tattare  Zigenare  Ziggis  Ziggis

B

Dickey-fuller Table

See [21] for the table with the significant values.

\[
\text{Pdf of } \tau = \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)
\]

Figure 18: Dickey-fuller distribution

36
References


