Predicting Linguistic Patterns Using ARIMA Models

Tim Isbister, Lisa Kaati

FOI, Uppsala University
Stockholm, Uppsala
Sweden, Sweden
tim.isbister@foi.se, lisa.kaati@it.uu.se

Abstract

In this work we explore whether it is possible to capture linguistic patterns and detect anomalous behavior of users on a web forum. In our experiments we use data from a Swedish web forum and investigate a forum specific linguistic pattern of 20 different users. If user data can be distinguished from white noise, statistical regression models such as ARIMA can be used to identify the underlying pattern and forecast data.

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 allows 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 artist, Lars Vilks (Conway, 2012).

Web forums are one place for extremist groups to spread their views and communicate. Monitoring web forum and groups is an important task for security agencies, researchers, and analysts. Manual inspection is one way to do this but since the number of forums and posts increase there is a growing need for more automated system that can assist humans in their analysis.

In this work we focus on time series analysis using autoregressive integrated moving average (ARIMA) models. The data that we use consist of a set of texts written by users on a Swedish xenophobic web forum `www.flashback.org/f226`, a sub-forum of the well-known Swedish forum `www.flashback.org`. The topics that are discussed are related to immigration and integration. The web forum started in 2007 and have more than 48000 members. The language of the web forum is Swedish and the analysis we have done uses a dictionary with forum specific words. The dictionary was created by doing a qualitative analysis of the web forum with a focus on words called ethnophaulisms, i.e., ethnic or racial slurs. The words on this list range from very crude and insulting ones (*skäggapa*), to words with ironically reversed meanings (*berikad*). The use of forum-specific jargon can be a way to express belonging to the group and therefore it is meaningful to detect such a jargon and to observe to what extent forum users uses the jargon over time. In particular new users and their usage of their adoption to the forum jargon are interesting to observe. In this work we use time series analysis to analyze and predict how individual forum users will use the forum specific jargon. The wordlist contains 80 words.

Using social media to predict future events has been done in many different ways. Some work has focused on predicting the outcomes of elections using social media. In (Anstead and O'Loughlin, 2015) the focus is on the relation between social media and public opinion and in (Anstead and O'Loughlin, 2013) several forms of social media is used in combination with ARIMA models to predict the outcome of the UK election 2010. However, to the best of our knowledge, ARIMA models have not been used to predict the linguistic pattern of individual social media users.

2. Predicting the Pattern

In this study we consider individuals usage of a forum specific jargon as a pattern over time. We use time series analysis to predict the pattern. The time series can be thought of as measuring a linguistic pattern in written text. When an adequate model is obtained for the series, the model can be viewed as a tool describing a possible pattern - in this case by an individual user. The model will be able to produce forecasts for the usage of forum specific jargon for the next post. 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. Single outliers will not be considered, rather unexpected bursts of activity, which might indicate that a author is becoming more fixated on a particular topic.

For each user we measure the usage of words from the forum specific jargon dictionary. All daily measurements are aggregated to a monthly value due to inactive users, and to provide the same distance between the measurements. To train the model, the first 7 years was used for training and the last 6 months up to 1 year as testing data.

3. Experiments and Results

Since time series modeling can be seen as a regression problem, we decided to use ARIMA models. ARIMA(p, d, q) is denoted as the autoregressive integrated

moving average model. Where p is the order of the auto regression, d represents the amount of differencing that needs to be done, if necessary, to achieve stationarity, finally q corresponds to the moving average of it's q past error terms. To select the most suitable model, we used Akakie information criteria combined with interpreting the correlograms (Jonathan D Cryer and Kung-Sik Chan, 2008). Measuring the forecast accuracy between different suggested models is made using the root mean squared error (Rob Hyndman, 2014).

4. Results

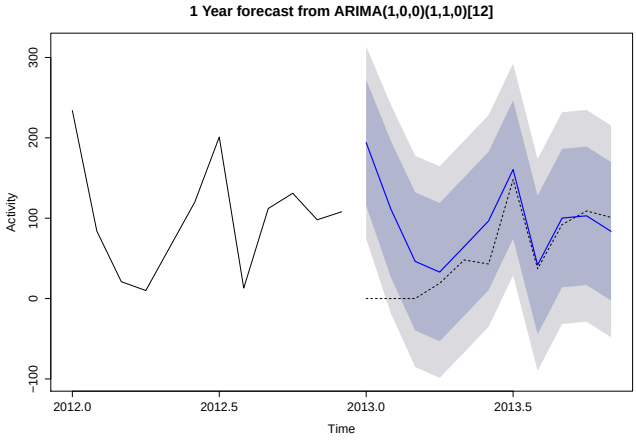

Figure 1: Time series and the prediction from User1

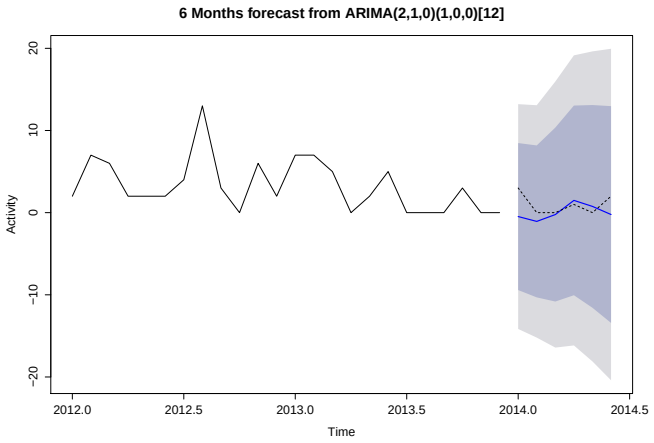

Figure 2: Time series and the prediction from User2

The ARIMA models manages to capture the predicted data within the 95% confidence intervals in 17 cases when making short term forecasts from 20 users. Figure 1 and Figure 2 shows the time series and the prediction for two users *User1* and *User2*. As can be seen, both Figure 1 and Figure 2 seems to capture the overall movement of the time series. *User1* was one of the most active users (when it comes to writing posts on the forum), with a significant amount of data compared to the others. In Figure 3, the the time series for *User3* is shown. The predicted data is outside the 95% confidence intervals, and therefore the linguistic pattern of the user could be seen as anomalous.

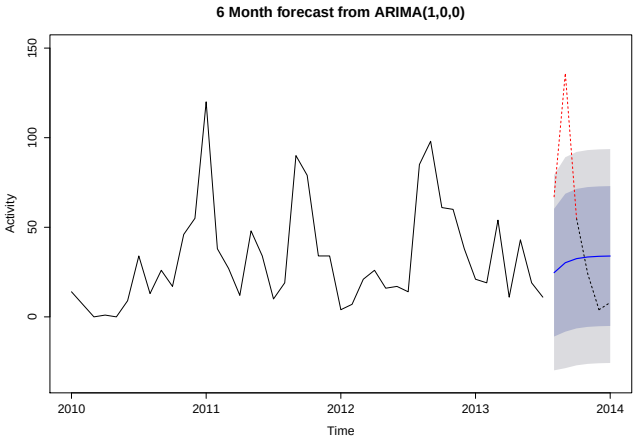

Figure 3: Time series and the prediction from User3

5. Discussion

The ARIMA models manages to capture the predicted data within the 95% confidence intervals in 17 cases when making short term forecasts from 20 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. If however, rapid burst of activity occurs as seen in Figure 3, it could indicate a change in the linguistic pattern, something that could be considered anomalous and perhaps something that should be investigated further.

References

M. Conway. From al-zarqawi to al-awlaki: The emergence of the internet as a new form of violent radical milieu *Vol. 2 No. 4, pp. 12-22.*

N. Anstead, B. O'Loughlin. 2015. *Social media analysis and public opinion: The 2010 uk general election*. Journal of Computer-Mediated Communication, Vol. 20, pp. 204-220, 2015.

F. Franch. 2013. *Wisdom of the crowds 2: 2010 uk election prediction with social media* Journal of Information Technology and Politics, Vol. 10, No. 1, pp. 57-71, 2013.

J. Cryer, K. Chan. 2008. *Time Series Analysis With Applications in R* Springer-Verlag New York Inc 2nd ed. 2008.

R. Hyndman. 2014. *Measuring forecast accuracy* http://www.robjhyndman.com/papers/forecast-accuracy.pdf.