Statistics Netherlands

Division of Process development, IT and Methodology
Methodology sector

Heerlen
The Netherlands

Social Media Sentiment and Consumer Confidence

Piet J.H. Daas and Marco J.H. Puts

Remarks:
The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. This paper is written for the workshop on using Big Data for forecasting and statistics, Frankfurt am Main, 7-8 April 2014.

Project number: -
BPA number: -
Date: *15 February 2014*

SOCIAL MEDIA SENTIMENT AND CONSUMER CONFIDENCE

Summary: Changes of the sentiment in Dutch public social media messages were compared with changes in monthly consumer confidence over a period of 3.5 years. This revealed that both were highly correlated (up to r = 0.9) and that both series cointegrated. This phenomenon is predominantly affected by changes in the sentiment of all Dutch public Facebook messages. Including various selections of public Twitter messages improved this association and improved the response to changes in sentiment. All Twitter messages or subsets thereof, without Facebook, never cointegrated. Changes in consumer confidence and social media sentiment were likely affected by an identical underlying phenomenon that more directly influenced consumer confidence. Social media picked up these changes with a short delay of around 7-days. In the paper the relation between social media sentiment and consumer confidence is discussed in depth.

Keywords: Social media, sentiment, Big Data, methodology, statistics, representativity

1. Introduction

Social media are used more and more by increasing numbers of people worldwide. In June 2013 eMarketing (2013) estimated that nearly a quarter of the world population is active on one or more social networks. This contribution, however, varies per country and region. In their 2012 autumn report Eurostat (2012) estimated that 42% of the people in the European Union above 12 years of age use social media at least once every week. But for some European countries this contribution is much higher and more frequent. In Iceland a staggering number of 77% use social media every day, followed by the Netherlands (60%; Stat. Neth., 2013b), Latvia (44%), Denmark and Sweden (both 43%). This makes social media in these countries a very interesting source for studies on social phenomena and other population related topics (Miller, 2011).

In recent years, quite a number of studies have been performed on the usability of social media messages. Although the majority seems to have a marketing perspective (Kaplan and Haenlien, 2010), some have looked at it from a more scientific point of view (Miller, 2011; Groves, 2011). Since we focus on the sentiment in social media in this paper, several important sentiment related studies are mentioned here. Landsdall-Welfare et al. (2012) used the sentiment in Twitter messages to nowcast the mood in the UK, whereas Bollen et al. (2010) and Rao and Srivastava (2012) attempted to predict the US stock market with Twitter sentiment. A considerable number of papers and reports have been written on this particular topic, for instance by O'Conner et al. (2010) that linked Twitter sentiment to the public opinion measured in several polls. This study also includes references to comparable studies by others. All studies claim to have succeeded fairly well in linking the overall sentiment in the specific social media platform studied with changes in the time series to which they were compared. This has resulted in several companies creating ‘rapid’ indicators based on social media, usually Twitter, for specific areas. The company Downside Hedge (2013) for instance uses Twitter sentiment for Stock Market analysis as a replacement for weekly surveys.

In this paper we focus on the sentiment in Dutch social media. This includes all publicly accessible messages on a considerable number of platforms, such as Twitter, Facebook and Linkedin, and also includes Dutch messages produced on websites, forums and in blogs. The sentiment in these messages is used as an indication for the overall sentiment in the Dutch population; e.g. the ‘mood’ of the Dutch nation. A first finding of this phenomenon has been presented at the 2013 New Techniques and Technologies for Statistics conference (Daas et al., 2013). In this paper we describe the relation between Dutch social media sentiment and consumer confidence and its potential use in depth. If changes in social media sentiment are indeed related to Dutch consumer confidence, the former could be used as a readily available indicator for changes in consumer confidence and, as such, may contribute or even provide an early indicator for an important official statistic. If such an indicator can be produced in a methodologically sound manner, these kind of statistics have the potential of being cheaper and faster than official statistics known to date.

2. Data and methods

2.1 Data sources

The study is based on two data sources. The first is consumer confidence data collected and determined by Statistics Netherlands. Consumer confidence is an index figure that indicates the extent to which households think that the economy is doing better or worse. The index is based on the sentiments of households on the economic climate in general and on their own financial situation (Stat. Neth., 2013a). During the first two weeks of each month Statistics Netherlands conducts the consumer confidence survey among around one thousand households. They are asked five questions that can be answered positively, negatively or neutrally; i.e. the situation has remained the same. The five questions are about the current and anticipated economic situation of the Netherlands, the current and anticipated financial situation of the household and if the current moment is considered a good time to buy large goods. The indicator for each question is calculated by subtracting the percentage of negative answers from the percentage of positive answers. Consumer confidence is the average net result of all five indicators. Findings for a particular month are reported in the week following the survey period; usually around the 20th of the month. Consumer confidence data is available in the electronic databank of Statistics Netherlands located at: http://statline.cbs.nl/.

Social media messages are the second data source used. Since large amounts of messages are created on various platforms, routinely collecting social media messages in large amounts is a tremendous effort. For our studies huge amounts of social media messages were needed on as many platforms as possible. We therefore purchased access to the collection of public social media messages gathered by the Dutch company Coosto (2014). This company routinely collects public social media messages written in the Dutch language on the most popular social media platforms in the country, such as Twitter, Facebook, Linkedin, Google+ and Hyves. Their data collection additionally includes Dutch messages and reactions posted on public blogs and forums and on many publically available web pages, such as those of newspapers and news sites. A total of 400.000 sources are continuously monitored. This has resulted in a collection composed of more than 3 billion messages covering the period of 2009 until the present. Around 2.5 million new messages are added per day. Through a secure online interface, the messages can be queried in a convenient fashion. Coosto also has a collection of social media messages produced in the UK.

2.2 Sentiment determination

Apart from the message's content and some basic information of the user, the sentiment of the messages collected is automatically determined by Coosto. This is done by checking whether a message expresses a negative or positive opinion. For this purpose a proprietary variant of a sentence-level based classification approach is used (for an overview see Pang and Lee, 2008). The approach strictly determines the overall sentiment of the combination of words included in each message. The sentiment classification of the words in the Dutch lexicon is used, in a fashion similar as described by van Assen et al. (2013), to which the sentiment of the informal words and emoticons used on social media are added (Velikovich et al., 2010). The overall sentiment of a message is assigned essentially as described by Esuli and Sebastiani (2006). This results in messages to which either a positive, negative or neutral label is assigned. Neutral messages exhibit no apparent sentiment, e.g. objective sentences. At the level of individual messages such a classification will obviously contain errors. However, since we are only interested in the aggregated sentiment of messages created during specific intervals (e.g. days, weeks, months), such errors will cancel out because of the enormous amounts of messages produced (see O'Conner et al., 2010 for more details). They may however still be potentially biased. Our studies usually included aggregates of 2 to 75 million messages per time interval studied. At the beginning of January 2013, Coosto adjusted their sentiment determination method by additionally assigning sentiment to messages containing smileys. This affected the average sentiment values of aggregates; they became more positive. To correct for this methodological change, the sentiment of daily aggregates in the two months before and after January 2013 were visually compared and aligned. Usually the difference was around 5%. Particularly for Facebook and Twitter messages, routine checks were performed to verify if and how this correction affected our findings by comparing the results obtained before and after January 2013.

2.3 Data selection and analysis

Via a secure web interface the database of collected public Dutch social media messages of Coosto was accessed. In the interface keywords, a time period and the various social media platforms to include were specified. Query results, such as the total number of messages and the number of positive and negative sentiment assigned messages included in the period studied, were exported at an aggregated level. Routinely, results were exported as daily aggregates in CSV-format for more rigorous analysis. For this the open source statistical software environment R was used (R Development Core Team, 2012). In R, the CSV-files were loaded and the total and number of positive and negative assigned sentiment messages were aggregated at selected time intervals, e.g. 7, 14, 21 or 28 days. The average sentiment for each interval was calculated by subtracting the percentage of positive classified messages from the percentage of negative classified messages included. Next, the social media sentiment findings were aligned with monthly consumer confidence data covering the same period.

The relation between series of individual and combinations of social media messages produced on various platforms and consumer confidence were compared with standard linear regression models. Models with and without interaction effects were considered. All messages produced during a specific time interval were aggregated and the development of the average sentiment was compared to consumer confidence. To determine the quality of the linear model for each series leave-out-one cross validation studies (Arlot and Celisse, 2010) were performed. Average correlation and cointegration values were determined (see below) and the average residual sum of squares was used as an additional measure of fit.

Pearson's product-moment correlation coefficients (*r*) of sentiment and consumer confidence were determined with the base cor function of R. Series were routinely checked by visual inspection; e.g. the creation of scatter plots. The concept of cointegration was used to check for stationary linear combinations of sentiment and consumer confidence (Murray, 1994). Cointegration was calculated according to the Engel-Granger two step method (Engle and Granger, 1987), i.e. after fitting a linear model an augmented Dickey–Fuller (adf) test was performed on the residuals. For this the adf-test function in the tseries package was used (Trapletti and Hornik, 2013). Series with a p-value below 0.05 were considered to cointegrate. Auto- and cross-covariance and -correlation of the residuals, to check for seasonality and trends, was studied with the appropriate functions in the astsa package (Stoffer, 2012). Independence of the residuals was also checked with the Durbin-Watson test in the lmtest package (Zeileis and Hothorn, 2002). Granger causality was used to study the predictive relation between social media sentiment and consumer confidence and vice-versa. These analysis rest on the assumption that if a particular variable affects another variable, changes in the first will systematically occur before changes in the other. If this is the case, lagged values of the first will exhibit a statistically significant correlation with the other variable. For these analyses the granger.test function in the lmtest package was used (Zeileis and Hothorn, 2002). Forecast skills scores were calculated as described by Murphy (1988) with 50% chance as the standard of reference and increase or decrease as possible outcomes.

3. Results

3.1 Exploratory analysis

Our initial studies revealed a sharp increase in the number of social media messages in the data set from June 2010 onwards. The latter corresponded to the starting period at which the collection of public Dutch Twitter messages was initiated at a large scale. After that around 75 million messages were added each month, corresponding to an average of 2.4 million messages a day. Because of this our studies focused on the period June 2010 until November 2013; a period of 42 months. Note that this is a relatively short period for time series analysis. An overview of the sentiment data, aggregated at a daily, weekly and monthly level, including an insert of the development of consumer confidence for the same period is shown in Figure 1. A visual comparison of the data in this figure suggests that consumer confidence and monthly aggregated social media sentiment display a similar development. The figure also reveals that daily sentiment fluctuated tremendously while weekly and monthly aggregates behaved much less volatile. Particularly prominent are the positive daily and weekly sentiment peaks near the end of December for 2010 and 2011. A similar situation has been reported for the UK (Lansdall–Welfare et al., 2012) and was due to an increase of more positive messages related to Christmas and New Year during that period. This was also the case here, but over the years the sentiment in the Dutch data set gradually decreased because of the increase in the number of negative messages complaining about firework nuisance. Studies focused on the identification of other patterns in the daily sentiment data suggested a weekly pattern, with a somewhat higher sentiment on Fridays and during the weekend. No clear other seasonal patterns became apparent. Hence, in subsequent studies it was decided to use aggregated sentiment data of one and more 7-day periods.

The platform dominating the social media data set is Twitter as 80% of all messages are composed of so-called ‘tweets’ (Table 1). Public Facebook messages comprise a bit more than 10% of the data set. In Table 1 an overview is given of the characteristics of messages created on all and on each of the various social media platforms discerned. The development of the average monthly sentiment of each of these sources is also compared to that of the original (non-seasonally adjusted) Dutch consumer confidence series covering the same period (Stat. Neth., 2013a). Both Pearson product moment correlation coefficients (*r*) and cointegration of the series are determined for all sources listed. Correlation is used to checks for a comparable development; the values before the slash sign in the last column of

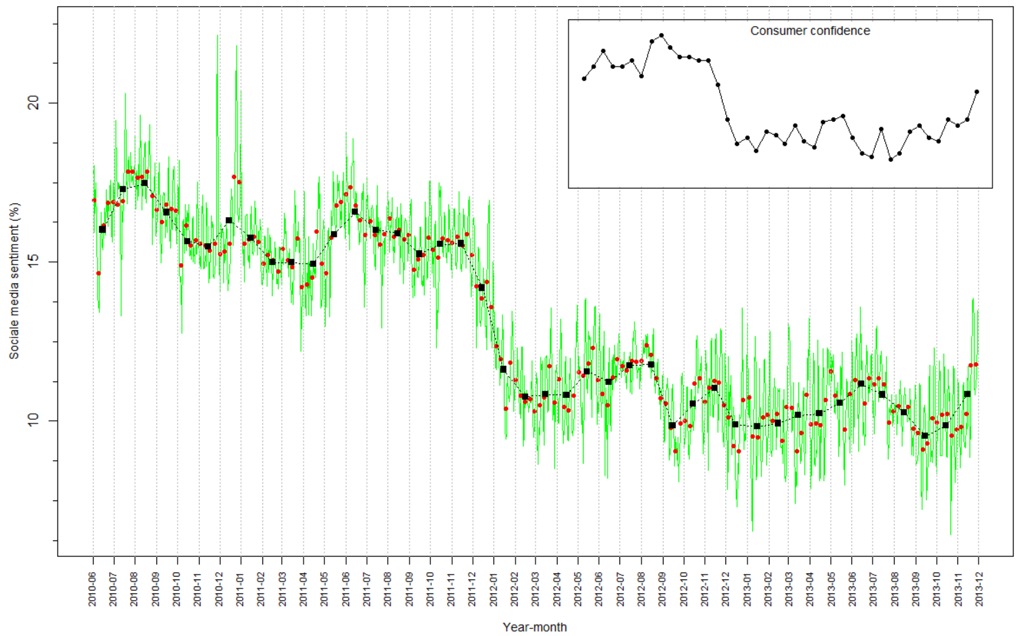

Figure 1. *Development of daily, weekly and monthly aggregates of social media sentiment from June 2010 until November 2013, in green, red and black, respectively. In the insert the development of consumer confidence is shown for the same period.*

Table 1. To reduce the risk of discovering spurious or false correlations, cointegration is additionally determined. Cointegration provides a stronger argument as it checks for a common stochastic drift, indicating that series exhibit fluctuations around a common trend (Engel and Granger, 1987). It is important to also consider the publication date of consumer confidence here. The survey is always conducted during the first 14 days of each month, and the figure –for that particular month– is published around the 20th of that month (Stat. Neth., 2013a). The exact date of the latter may fluctuate a few days depending on the relative position of the working days following the survey period. Monthly sentiment aggregates will therefore also include days in which consumer confidence –for that particular month– is already publically known. We therefore also compared the 'monthly' sentiment of the combination of social media messages produced in the second half of the previous month (after the survey period for that month) and those produced in the first half of the current month (during the survey period for that month and *before* the moment of publication). The messages produced during such periods make maximum use of the sentiment-related information available and are expected to be less likely influenced by the consumer confidence findings for the month to which they are compared. The results for this adjusted time interval are shown in the last column of Table 1 (after the slash sign).

These studies reveal that changes in the sentiment of public Facebook messages not only highly correlate with consumer confidence, r = 0.81 and 0.85 depending on the time interval, but that these series also cointegrate. This clearly demonstrates a good *association* between both series, but says nothing about an underlying cause. Sentiment data of the other platforms and of the combination of all messages display various degrees of correlations that do *not* cointegrate, suggesting none or a much poorer association. To test whether changes in Facebook sentiment preceded changes in consumer confidence or vice-versa, Granger causality analysis were performed. These analysis rest on the assumption that if a particular variable affects another variable, changes in the first will systematically occur before changes in the other. If this is the case, lagged values of the first will exhibit a statistically significant correlation with the other variable. Lagging the period for Facebook on consumer confidence by one or more months did not reveal an additional effect. Lagging the period for consumer confidence data by one month, however, did reveal a significant effect on Facebook sentiment; $p < 0.001$ for the series in which the days in each month completely coincided and $p < 0.05$ for the series with an adjusted time interval. This and the difference between those p-values suggest that Facebook sentiment is very likely affected by consumer confidence. Since this is even the case for the series with the adjusted time interval, which did *not* include messages produced during the time consumer confidence is published for that particular month, this suggests that both might be affected by a common underlying cause.

Visual plots of the relative development of daily aggregated sentiment in each of the social media platforms provided additional insights. In Figure 2, the development of the sentiment in Twitter clearly demonstrates a decrease around the period consumer confidence dropped (compare these with the insert in Figure 1) and it is the only

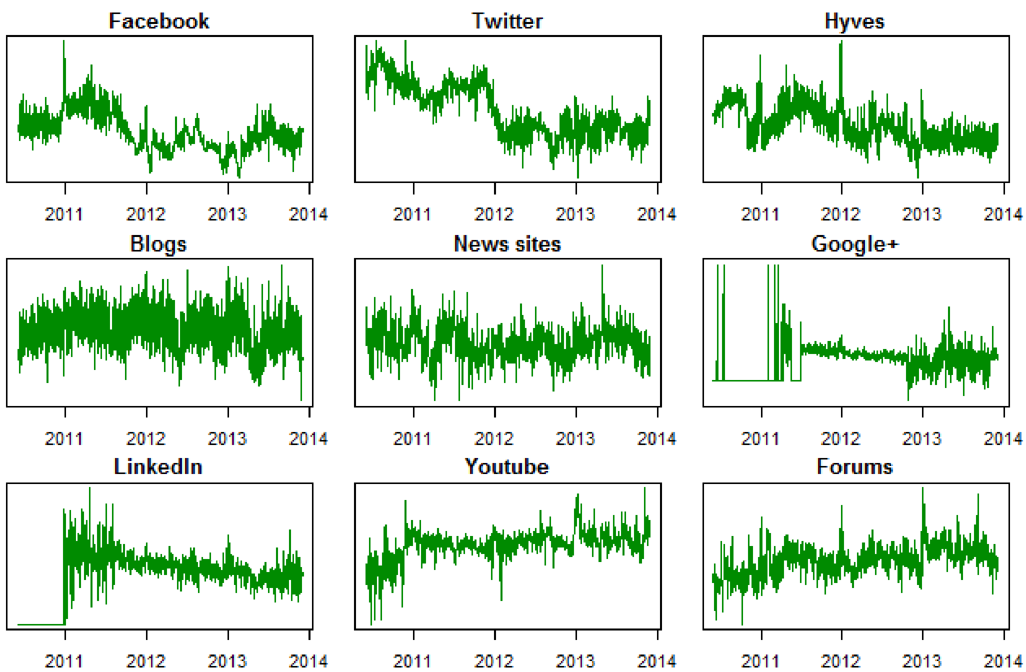

Figure 2. Relative development of daily aggregates of social media sentiment on the various platforms during the period studied. Facebook and Twitter account for 10 and 80% of all messages, respectively.

platform with a clear increase in sentiment at the end of the series; i.e. November 2013. This is also the onset of a steady increase in consumer confidence in the Netherlands (Stat. Neth., 2013a). This prompted us to investigate if any combination or subset(s) of social media messages were able to capture these phenomena.

3.2 Effect of Facebook and Twitter messages

Since Twitter and Facebook clearly dominated the social media data set, the collected messages on these platforms were investigated thoroughly. Messages produced on the other platforms were also included in our studies but provided no additional information and are therefore no longer discussed. This is not unexpected considering the small amounts of messages they contain (Table 1). Many combinations of messages produced on Facebook and/or Twitter with and without using specific words as selection criteria were tested. The latter approach was inspired by an earlier study performed at our office that revealed that nearly 50% of all Twitter messages produced in the Netherlands can be considered 'pointless-babble' (Daas et al. 2012). The latter property made them potentially less interesting. Perhaps, selecting messages only containing specific words could positively affect the association between social media sentiment and consumer confidence. The relation between series of individual and combinations of Facebook and Twitter messages, with and without using specific words as selection criteria, and consumer confidence were compared with standard linear regression models. Models with and without interaction effects were considered.

Messages were aggregated and assigned to months according to the adjusted time interval described above. All messages produced during a specific time interval were aggregated and the development of the average sentiment was compared to consumer confidence. Leave-out-one cross validation studies (Arlot and Celisse, 2010) were used to determine the quality of the model. Average correlation and cointegration values were determined and the average residual sum of squares was used as an additional measure of fit. Granger causality of the effect of the sentiment on consumer confidence and vice-versa were also determined. This work revealed that a considerable number of the combination of *all* public Facebook messages and specific word selected Twitter messages displayed high correlation coefficients with consumer confidence. Many of these series also cointegrated. Combinations in which selections of Facebook messages were included always performed worse. The effect of including interaction effects in the models when using combinations of sources varied. In Table 2 the results of the best performing combinations of all Facebook and selected Twitter messages are shown. They all display high correlation coefficients (at least $r = 0.86$) and they all cointegrate. The findings for Facebook alone are listed at the top of the table, followed by the combination of Facebook and all Twitter messages. In Table 2, three distinct types of selection criteria are shown. The first is the use of words equal or related to consumer confidence or to the wordings used in the questions asked to determine it; no. 3-6 in Table 2. The second approach focuses on messages containing personal pronouns, such as 'I', 'me', 'you' and 'us', reflecting personal or group experiences; no. 7-9.

The third group contains words or combinations of words used most often in the Dutch language, both written and spoken; no. 10-14. Examples of this are the Dutch articles and the words ‘this’ and ‘that’. In addition, combinations of words included in the second and third group are also considered, no. 15-20.

The regression models without interaction indicated that in all cases shown in Table 2, the inclusion of both Facebook and Twitter messages significantly contributed to the model ($p < 0.001$) and that the β-coefficients of both sources were positive. With sentiment on the x-axis and confidence on the y-axis, the β_0-coefficient was negative in all cases, reflecting the fact that the average social media sentiment is much more positive compared to the average consumer confidence over the period studied (see also Table 1). Visual inspection and checks for autocorrelation of the residuals of the models revealed no apparent trend. The results of Granger causality analysis, to specifically check if the sentiment in any of the combination of sources preceded consumer confidence, differed greatly for the various combinations shown and for the models with and without interaction (Table 2). Usually the p-values for the models including an interaction component were lower. Several of the combinations listed in Table 2 revealed a potential effect of sentiment on consumer confidence; $p < 0.01$. The latter suggests a probability for a preceding effect of social media sentiment on consumer confidence. These findings support the idea that some of the combinations listed in Table 2 could be able to detect ‘upcoming’ changes in consumer confidence. What Table 2 also reveals is that each of the different types of selection criteria tested works. Use of specific consumer confidence or economy related words provides positive results (no. 5 and 6 in Table 2) as do all of the personal pronouns (no. 7-9) or the inclusion of words used with a high frequency in the Dutch language (no. 10-13). The most intriguing result in Table 2 is the fact that after the combined top 10 of the most frequently written and spoken Dutch words are used as selection criteria for Twitter messages (no. 10 in Table 2) merely 65% of the total number of messages is included. Inspecting the excluded messages revealed that more than 90% of them can be designated as pointless babble.

3.3 Effect of different 7-day periods and forecasting properties

Since the consumer confidence survey is conducted during the first half of the month, attention was also put into comparing the average social media sentiment of various 7-day periods before, during and after the survey period. These periods started 14 days before and ended 28 days after the beginning of the month. All combinations listed in Table 2 were tested. For each series leave-out-one cross validation studies were performed. This revealed that very high correlation coefficients, up to $r = 0.93$, were found for the period coinciding with the second half of the survey period; i.e. day 8 until 14 (Table 3). Every combination in this –and in the subsequent two 7-day periods– cointegrated. Correlations for the first 7 days of the week, i.e. the first part of the survey period, were somewhat lower and did *not* cointegrate for any combination. The latter suggests a clear distinction between the development of the sentiment of social media messages produced during the first and second half of the survey period. In this respect it is important to note that routinely the response to the survey is highest during the first 7-days in which usually around 70% of the total response is obtained. These findings are clearly *not* reflected in the sentiment of the social media messages produced during that period. Best overall performing model, based on correlation, cointegration and the residual sum of squares, for any of the 7-day periods listed in Table 3, was the combination of all Facebook and Twitter messages containing any of the Dutch articles, that or personal pronouns (no. 20). When aggregates of longer periods were compared, e.g. 14, 21 and 28-days, all 28-day aggregates and any aggregate covering day 8 until 21 or day 15 until 28 cointegrated with consumer confidence. Best 28-day period was the combinations of messages produced in the last 14-days of the previous month and the first 14-days of the current month. These results are shown in Table 2. Here, again the best results were obtained for the combination of all Facebook and Twitter messages containing any of the Dutch articles, that or personal pronouns (no. 20).

The combinations were also checked for their ability to pick up the increase in consumer confidence observed in November 2013 (Figure 1). Sentiment combinations were fitted to consumer confidence data with the exception of the last month. Messages produced during various 7-day intervals, as shown in Table 3, or combinations thereof were tested and the predicted value of month 42 was compared with the actual increase measured. Linear models with and without interaction were used. The best performing models used the sentiment of messages produced during the 22-28th day of the month, all other periods performed much worse. Models without interaction usually performed somewhat better as did combinations including large amounts of Twitter messages. The combination of all Facebook and all Twitter messages performed best; 93% of the increase was picked-up. Next was the combination of Facebook and Twitter messages containing the Dutch articles, that and personal pronouns (no. 20) with an increase of 88%. Attempts to use changes in social media sentiment to *predict* changes in consumer confidence with any of the combinations and periods listed in Table 3 were unsuccessful. Here, for each month an increase or decrease in consumer confidence was predicted and 50% chance was used as the reference forecast, i.e. a forecast skill score of zero. A maximum score (Murphy, 1988) of 0.12 was found for a number of the combination shown in Table 3 for messages produced during the 22-28th day of the month, where a value of one identifies the perfect score.

3.4 Comparison with UK-data

At the end of the study the relation between the sentiment in social media messages and consumer confidence was checked for UK-data. The Dutch firm that provided access to Dutch messages also routinely collects public social media messages produced on various platforms in the UK. Again data were available from June 2010 onwards. Results were compared to the monthly consumer confidence barometer of GfK (2014). The latter displayed a more volatile behaviour reflecting a somewhat poorer quality compared to the Statistics Netherlands survey results. Even with this in mind, it was found that social media sentiment in publicly available social media messages in the UK correlated highly (r = 0.8) with UK consumer confidence. These results were, however, only achieved if one specific month, August 2012, was removed from the sentiment series. This period roughly covered the Olympic Games held in London from July 27 until August 12. During these games social media was very actively used by athletes, journalists, the Olympic committee and the public to inform and cheer-up the Olympic athletes (see SportLaw, 2012 and references therein). Especially when an athlete won gold, positive sentiment peaked tremendously. This change in routine use of social media, by UK and non-UK inhabitants, clearly negatively affected the more common relation observed –before and after the Olympic Games– between social media sentiment and consumer confidence. With the understandable exception of August 2012, the UK-results corroborate the Dutch findings.

4. Discussion

The results described above confirm that there is an association between (changes in) social media sentiment and consumer confidence for both the Netherlands and the UK. This relation remained stable during the period investigated in our studies; with the exception of August 2012 in the UK. This indicated that major changes in the behaviour of the public on social media, such as those caused by major events (e.g. the Olympic Games), can have a disturbing effect. Studies of Dutch social media indicated that public Facebook messages alone are already capable of capturing this phenomenon. This is interesting as the majority of the Dutch population active on social media, about 70%, report they use Facebook (Stat. Neth., 2013b). Granger causality studies for Facebook demonstrated that the changes in the sentiment in the public messages produced on this platform are very likely affected *after* consumer confidence changes. The combination of public Facebook messages and Twitter messages containing any of the Dutch articles, that or personal pronouns (no. 20 in Table 2 and 3) are best at capturing this relation and responded better to *changes* in sentiment. The fact that Twitter messages containing such generally used words are –in combination with Facebook– most effective suggests that quite a general ‘mood’ is (indirectly) measured. Since models including an interaction component had a tendency to perform somewhat better this additionally supports the idea that a general occurring ‘mood’ is measured as this suggests that sentiment changes occurring on both platforms are considered additionally important. An explanation for the phenomenon observed can be found in the Appraisal-Tendency Framework (Han et al. 2007), which is concerned with consumer decision making. In this framework it is claimed that a consumer decision is influenced by two kinds of emotions, i.e. the *incidental* emotion and the *integral* emotion. Here, the incidental emotion is irrelevant for a decision at stake whereas the integral emotion is relevant. Based on this theory, consumer confidence is likely to be mainly influenced by the incidental emotion, as consumer confidence is not measured in relation to an actual decision to buy something. With this in mind our research findings suggest that the sentiment in social media messages might reflect the incidental emotion in the part of the population active on social media. Because of the general nature of the latter, it would not surprise us if someone would denote it the ‘mood’ of the nation (Lansdall-Welfare et al., 2012).

Comparing messages produced during different 7-day periods revealed that the sentiment in messages produced on days 8 until 14 of each month correlated best with consumer confidence; r = 0.93 were the highest correlations found. Cointegration indicated that both series share a common stochastic drift, supporting the idea of long-term stability. Since the response to the consumer confidence survey is predominantly obtained in the first 7-days of the 14-day survey period, which starts at the beginning of the month, this supports the idea of a (short) delay between changes in consumer confidence and social media sentiment. A delay was also observed when only Facebook messages were compared. This, combined with the fact that attempts to predict consumer confidence with social media sentiment performed very poor, additionally supports the notion of a lag between both. All our results are consistent with the notion that changes in an apparent ‘mood’ of the Dutch population *both* affect the part of the population responding to the consumer confidence survey and the part of the population creating public Facebook and Twitter messages in the same direction and, for social media, with a lag of around 7 days. As such, developing a ‘real-time’ indicator of the ‘mood’ of the nation based on social media seems only possible if one accepts a short, possibly 7-day, lag. Because of this lag, the claim of Lansdall-Welfare et al. (2012) that social media could be used to ‘nowcast ’ the mood of the nation is not fully supported by our findings. It is however still a bit faster and can certainly be determined more frequently than the survey that probably also reflects this ‘mood’ best; the consumer confidence survey. Combining both would be ideal.

It was also found that the sentiment in social media is biased as it is much more positive then consumer confidence; if one assumes that the latter is closer to the truth. Based on the notion that a similar ‘mood’ affects both, clearly the public active on social media has a tendency to respond much more positive. Perhaps this is a reflection of the tendency of people active on social media to –more likely– report the positive things occurring in their life. It could also result from a difference in the age composition of the persons included in both sources as younger and elderly people have a tendency to respond more and less positive, respectively (Stat. Neth., 2013c). The observed bias could also be an indication that two –apparently cointegrated– non-identical phenomena are compared. Quantifying the contributing effect of social media sentiment on consumer confidence, for various age groups, could be a way to determine which of these options is more likely to be correct.

Comparing the people responding to the survey and those active on social media brings us to the most intriguing part of this study: the relation between consumer confidence and social media sentiment from a population point of view. Even though our study revealed that there clearly is an association between both, the units used to determine confidence and sentiment are obviously different. The units of the consumer confidence survey are households, from which a representative (the head; a person) is contacted and interviewed (Stat. Neth., 2013a). Usually around a 1000 persons respond to the survey each month. For social media, the public messages written in Dutch are collected and treaded as if they are the units. Of these, millions to tens of millions are produced per month and included in our analysis. Since confidence and sentiment are calculated in exactly the same way, the percentage of positives minus the percentage of negatives, their development can be easily compared. Based on this and the fact that the population involved in the consumer confidence survey is a representative part of the Dutch population (Stat. Neth., 2013b), one can only conclude that (the changes in) both variables (sentiment and confidence) must also be representative (Buelens et al., 2013). Is this, in the case of social media, an example of the law of the large numbers in action? One is tempted to conclude this as the collection of social media messages is clearly an example of observational data; data collected without a design. From a Big Data perspective (Daas and Puts, 2014) it is good in such cases to strive to completely cover such a dataset. And this is exactly what the Dutch company does; it attempts to collect as many public social media messages on as many publically accessible platforms as possible. Another explanation could be that the underlying phenomenon studied is simply less affected by differences in the composition of the population from which each variable is derived. As a result, even despite these differences the changes observed for both sentiment and confidence are expected to behave quite similar. For monthly and weekly aggregated sentiment data this might be the case, but for daily sentiment this obviously is not (see Figure 1). In addition, the fact that younger people are more active on social media (Eurostat, 2012; Stat. Neth., 2013b) and respond more positive in the consumer confidence survey (Stat. Neth., 2013c) could explain the bias observed but it does not support the idea of a phenomenon poorly affected by variations in the composition of the population included.

Clearly more scientific research is needed to fully comprehend the phenomenon described in this study. Providing no major events occur in the Netherlands that affect the behaviour on social media and assuming that the bias between both series remains constant, our findings support the idea that social media could be used to enhance official statistics. For example by producing a weekly confidence indicator based on social media sentiment. First results reveal that this is a much more volatile figure. In this respect it is also interesting to investigate the relation to the five individual indicators on which consumer confidence is based. In addition, one could also attempt to 'extract' opinions on other topics from social media, providing enough messages are available that capture the topic of interest. Future studies will focus on all of the above.

Acknowledgments

The authors gratefully acknowledge their Statistical Netherlands colleagues Bart Buelens, Jan van den Brakel, Floris van Ruth, Joep Burger, Barteld Braaksma and Peter Struijs for stimulating discussions and constructive remarks. Thierry Thijsen and Frank Scheelen from Coosto are gratefully acknowledged for their assistance during and feedback on the studies described in this paper. This work would not have been possible without the support from the Innovation program at Statistics Netherlands.

References

Arlot, S., Celisse, A. (2010) A survey of cross-validation procedures for model selection. *Statistics Surveys* 4, 40–79.

Bollen J., Mao H., Zeng, X-J. (2011) Twitter mood predicts the stock market. *Journal of Computational Science* 2 (1), 1-8.

Buelens, B., Daas, P., Burger, J., Puts, M., van den Brakel, J. (2013) Selectivity of Big Data. Internal report, Statistics Netherlands, Heerlen, The Netherlands.

Coosto (2014) Online Radar website. Located at: http://www.coosto.com/

Daas, P.J.H., Puts, M.J.H. (2014) Big Data as a Source of Statistical Information. *The Survey Statistician* 69, 22-31.

Daas, P.J.H., Puts, M.J., Buelens, B., van den Hurk, P.A.M. (2013) Big Data and Official Statistics. *Paper for the 2013 New Techniques and Technologies for Statistics conference*, Brussels, Belgium.

Daas, P.J.H., Roos, M., van de Ven, M., Neroni, J. (2012) Twitter as a potential data source for statistics. *Discussion paper* 201221, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Downside Hedge (2013) Twitter indicators for Stock Market analysis. Located at: http://www.downsidehedge.com/twitter-indicators/

eMarketeers (2013) Social Networking Reaches Nearly One in Four Around the World. Located at: http://www.emarketer.com/Article/Social-Networking-Reaches-Nearly-One-Four-Around-World/1009976

Engle, R.F., Granger, C.W.J. (1987) Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica* 55 (2), 251-276.

Esuli, A., Sebastiani, F. (2006) Senti WordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion mining. *Proceedings of Language Resources and Evaluation (LREC),* pp. 2200-2004. Located at: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

Eurostat (2012) Media use in the European Union. *Standard Eurobarometer report* 78, autumn. Located at: http://ec.europa.eu/public_opinion/archives/eb/eb78/eb78_media_en.pdf

GfK (2014) Consumer confidence barometer data. Located at: http://www.gfk.com/uk/Documents/Press-Releases/Charts%20%28December%202013 %29%20 %282%29.pdf

Groves, R.M. (2011) Three Eras of Survey Research. *Public Opinion Quarterly* 75 (5), 861-871.

Han, S., Lerner, J.S., Keltner, D. (2007) Feelings and Consumer Decision Making: The Appraisal-Tendency Framework. *Journal of Consumer Psychology* 17 (3), 158–168.

Kaplan A.M., Haenlien, M. (2010) Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons* 53 (1), 59–68.

Lansdall-Welfare, T., Lampos, V., Cristianini, N. (2012) Nowcasting the mood of the nation. *Significance* 9 (4), 26-28.

Miller, G. (2011) Social Scientists Wade Into the Tweet Stream. *Science* 333 (6051), 1814-1815.

Murphy, A.H. (1988) Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient. *Monthly Weather Review* 116 (12), 2417–2424.

Murray, M.P. (1994) A Drunk and Her Dog: An Illustration of Cointegration and Error Correction. *The American Statistician* 48 (1), 37-39.

O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A. (2010) From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, May 23-26, Washington DC, USA.

Pang and Lee (2008) Opinion and sentiment mining. *Foundations and Trends in Information Retrieval* 2 (1-2), 1-135.

R Development Core Team. (2009). R:A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Rao, T., Srivastava, S. (2012) Twitter Sentiment Analysis: How To Hedge Your Bets In The Stock Markets. Located at: http://arxiv.org/pdf/1212.1107.pdf

SportLaw (2012) Socialympics: How Sports Organizations and Athletes used Social Media at London 2012. Located at: http://www.sportlaw.ca/wp-content/uploads/2013/01/Social-Media-and-the-Games.pdf

Stat. Neth. (2013a) Consumer confidence, Web page of Statistics Netherlands. Located at: http://www.cbs.nl/en-GB/menu/themas/dossiers/conjunctuur/publicaties/conjunctuurbericht/inhoud/conjunctuurklok/toelichtingen/ck-03.htm

Stat. Neth. (2013b) Seven in ten internet users active on social media. Statistics Netherlands web magazine, October 4. Located at: http://www.cbs.nl/en-GB/menu/themas/vrije-tijd-cultuur/publicaties/artikelen/archief/2013/2013-3907-wm.htm

Stat. Neth. (2013c) Consumer confidence according to age groups (in Dutch). Statline pages of Statistics Netherlands. Located at: http://statline.cbs.nl/StatWeb/publication/?DM=SLNL&PA=71698ned&D1=0-5&D2=a&D3=0,3-5&D4=54-59&HDR=T,G1&STB=G3,G2&VW=T

Stoffer, D. (2012). astsa: Applied Statistical Time Series Analysis. R package version 1.1.

Trapletti, A., Hornik, K. (2013). tseries: Time Series Analysis and Computational Finance. R package version 0.10-32.

Van Assem, M., Isaac, A., van Ossenbruggen, J. (2013) Wordnet 3.0. Located at: http://datahub.io/nl/dataset/vu-wordnet

Velikovich, L.; Blair-Goldensohn, S.; Hannan, K.; and Mc-Donald, R. (2010) The viability of web-dervied polarity lexicons. *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 777–785.

Zeileis,A., Hothorn, T. (2002) Diagnostic Checking in Regression Relationships. *R News* 2 (3), 7-10.

Tables

Table 1. Sociale media message properties for various platforms and their correlation with consumer confidence

Social media platform	Number of social media messages[1]	Number of messages as percentage of total (%)	Number of sentiment assigned messages of total (%)	Average sentiment (%)	Correlation coefficient of monthly sentiment index and consumer confidence (*r*)[2]
All platforms combined	3,161,538,534	100	36.8	13.0	0.75 / 0.78
Facebook	334,894,060	10.6	34.1	20.5	0.81* / 0.85*
Twitter	2,531,627,287	80.1	35.7	11.6	0.68 / 0.70
Hyves	45,288,698	1.4	45.2	22.8	0.50 / 0.58
News sites	56,598,982	1.8	46.1	-1.5	0.37 / 0.26
Blogs	49,366,962	1.6	64.4	39.4	0.25 / 0.22
Google+	644,042	0.02	39.8	19.6	-0.04 / -0.09
Linkedin	565,811	0.02	47.9	25.7	-0.23 / -0.25
Youtube	5,665,644	0.2	43.6	16.7	-0.37 / -0.41
Forums	136,887,048	4.3	47.3	15.9	-0.45 / -0.49

[1] period covered June 2010 untill November 2013

[2] values after the slash cover messages produced in second half of previous month and first half of current month (see text)

[*] cointegration

Table 2. *Social media message properties for combinations of Facebook and Twitter messages and their correlation with consumer confidence*

	Effect of Twitter messages containing specific words in combination with all Facebook messages	Number of messages[1]	Number of messages as percentage of total messages collected (%)	Number of messages as percentage of total Twitter messages collected (%)	Average sentiment (%)	Correlation coefficient (*r*) without / with interaction[3]	Sum of squares without / with interaction[3]	Granger causality p-value; effect of sentiment on confidence[3]	Granger causality p-value; effect of confidence on sentiment[3]
1	No Twitter messages (Facebook alone)	334,854,088	10.6	-	20.3	0.85*	1676	0.0934	0.0351
2	All Twitter messages	2,861,335,567	90.7	100	12.4	0.87* / 0.89*	1468 / 1286	0.0720 / 0.0361	0.0097 / 0.0024
3	consumer, confidence	334,881,109	10.6	0.001	20.3	0.86* / 0.86*	1629 / 1623	0.0715 / 0.1127	0.0762 / 0.0741
4	consumer, confidence, economy, finance, spending	339,076,031	10.8	0.2	19.8	0.88* / 0.88*	1456 / 1452	0.0326 / 0.0400	0.0119 / 0.0142
5	consumer, confidence, economy, finance, spending & synoniems there of	361,217,766	11.5	1.0	18.9	0.89* / 0.89*	1289 / 1265	0.0122 / 0.0044	0.0732 / 0.0535
6	economy, job, jobs	339,774,637	10.8	0.2	20.0	0.88* / 0.88*	1408 / 1412	0.0092 / 0.0099	0.0373 / 0.0401
7	I	848,063,303	26.9	20.3	12.5	0.89* / 0.90*	1263 / 1157	0.0419 / 0.0073	0.0215 / 0.0142
8	I, me	990,225,039	31.4	25.9	11.5	0.90* / 0.91*	1240 / 1125	0.0450 / 0.0078	0.0162 / 0.0112
9	I, me, you, we, he, she & other personal pronouns	1,295,209,897	41.1	38.0	11.8	0.90* / 0.91*	1159 / 1080	0.0278 / 0.0051	0.0227 / 0.0171
10	Combination of top 10 of most frequently spoken and written Dutch words (the, from, a/an, and, in, is, that, this, with, are, yes, I, but, not, you, to, on, for, uh)	1,976,214,034	62.7	65.0	12.5	0.89* / 0.90*	1252 / 1150	0.0224 / 0.0046	0.0184 / 0.0120
11	the (in Dutch: de)[2]	672,233,894	21.3	13.4	15.3	0.89* / 0.90*	1260 / 1210	0.0448 / 0.0143	0.0198 / 0.0103
12	the (in Dutch: het)[2]	585,047,246	18.6	9.9	15.8	0.89* / 0.90*	1264 / 1212	0.0266 / 0.0088	0.0323 / 0.0205
13	a/an (in Dutch: een)	627,740,894	19.9	11.6	16.6	0.90* / 0.90*	1232 / 1226	0.0074 / 0.0041	0.0126 / 0.0109
14	that (in Dutch: dat)	545,575,408	17.3	8.3	15.5	0.89* / 0.90*	1275 / 1205	0.0524 / 0.0119	0.0234 / 0.0187
15	the, a/an (Dutch articles)	1,059,162,973	33,6	28.7	14,5	0.90* / 0.90*	1232 / 1198	0.0180 / 0.0063	0.0214 / 0.0147
16	the (het), I	1,026,797,559	32.6	27.4	12.5	0.90* / 0.90*	1241 / 1157	0.0312 / 0.0061	0.0257 / 0.0174
17	a/an, I	1,062,599,886	33.7	28.8	12.9	0.90* / 0.90*	1206 / 1148	0.0202 / 0.0044	0.0208 / 0.0154
18	the, a/an, I	1,383,657,115	43.9	41.5	12.8	0.90* / 0.90*	1220 / 1162	0.0222 / 0.0051	0.0229 / 0.0159
19	the, a/an, I, that	1,446,902,927	45.9	44.0	12.7	0.90* / 0.90*	1223 / 1162	0.0251 / 0.0056	0.0226 / 0.0165
20	the, a/an, that, I, me, you, we, he, she & other personal pronouns	1,711,886,042	54.3	54.5	12.2	0.90* / 0.91*	1179 / 1114	0.0215 / 0.0045	0.0215 / 0.0160

[1] period covered June 2010 untill November 2013
[2] Dutch has two definite articles: 'de' and 'het'
[3] Average results of 42 leave-one-out cross validations
* cointegration

Table 3. *Social media message properties for combinations of Facebook and Twitter messages produced during various 7-day periods and their correlation to consumer confidence*[1]

	Effect of Twitter messages containing specific words in combination with all Facebook messages	Period -2 Day -14 until -8 Previous month Correlation coëf (*r*) (residual sum of sqrs)	Period -1 Day -7 until -1 Previous month Correlation coëf (*r*) (residual sum of sqrs)	Period 1 Day 1 until 7 Current month Correlation coëf (*r*) (residual sum of sqrs)	Period 2 Day 8 until 14 Current month Correlation coëf (*r*) (residual sum of sqrs)	Period 3 Day 15 until 21 Current month Correlation coëf (*r*) (residual sum of sqrs)	Period 4 Day 22 until 28 Current month Correlation coëf (*r*) (residual sum of sqrs)
1	No Twitter messages (Facebook alone)	0.79 (2322)	0.78 (2411)	0.79 (2263)	0.82* (1951)	0.79* (2251)	0.85* (2662)
2	All Twitter messages	0.82 (2031)	0.84 (1815)	0.84 (1890)	0.89* (1316)	0.87* (1576)	0.81* (2210)
3	consumer, confidence	0.47 (2339)	0.79* (2321)	0.79 (2318)	0.83* (1922)	0.80* (2254)	0.76* (2642)
4	consumer, confidence, economy, finance, spending	0.81 (2149)	0.84* (1833)	0.81 (2156)	0.87* (1548)	0.85* (1785)	0.78* (2479)
5	consumer, confidence, economy, finance, spending & synoniems there of	0.86* (1575)	0.87 (1485)	0.87 (1480)	0.91* (1128)	0.87* (1497)	0.83* (1945)
6	economy, job, jobs	0.84 (1862)	0.85 (1177)	0.85 (1714)	0.89* (1288)	0.85* (1779)	0.81* (2231)
7	I	0.85 (1717)	0.87 (1557)	0.89 (1343)	0.92* (922)	0.89* 1347)	0.84* (1826)
8	I, me	0.85 (1757)	0.86 (1625)	0.89 (1344)	0.92* (924)	0.88* (1358)	0.85* (1788)
9	I, me, you, we, he, she & other personal pronouns	0.86* (1664)	0.87 (1495)	0.90 (1209)	0.93* (875)	0.89* (1294)	0.85* (1719)
10	Combination of top 10 of most frequently spoken and written Dutch words (the, from, a/an, and, in, is, that, this, with, are, yes, I, but, not, you, to, on, for, uh)	0.84 (1794)	0.87 (1480)	0.87 (1486)	0.91* (1061)	0.88* (1367)	0.73 (2965)
11	the (in Dutch: de)[2]	0.83 (1970)	0.88 (1454)	0.88 (1455)	0.92* (986)	0.89* (1344)	0.81* (2145)
12	the (in Dutch: het)[2]	0.85 (1732)	0.88 (1424)	0.88 (1388)	0.92* (933)	0.88* (1393)	0.83* (1947)
13	a/an (in Dutch: een)	0.83 (1945)	0.88* (1461)	0.90 (1246)	0.93* (850)	0.89* (1256)	0.81* (2155)
14	that	0.86 (1632)	0.88 (1412)	0.88 (1366)	0.92* (934)	0.88* (1366)	0.83* (1957)
15	the, a/an (Dutch articles)	0.84 (1825)	0.88 (1385)	0.89 (1290)	0.93* (901)	0.89* (1311)	0.82* (2038)
16	the (het), I	0.85 (1727)	0.88 (1463)	0.89 (1304)	0.93* (899)	0.89* (1347)	0.84* (1834)
17	a/an, I	0.84 (1835)	0.87* (1525)	0.89 (1247)	0.93* (858)	0.89* (1292)	0.84* (1880)
18	the, a/an, I	0.85 (1787)	0.88 (1426)	0.89 (1263)	0.93* (879)	0.89* (1320)	0.83* (1934)
19	the, a/an, I, that	0.85 (1767)	0.88 (1425)	0.89 (1269)	0.93* (881)	0.89* (1322)	0.83* (1937)
20	the, a/an, that, I, me, you, we, he, she & other personal pronouns	0.85 (1729)	0.88 (1435)	0.90 (1225)	0.93* (869)	0.89* (1302)	0.84* (1854)

[1] Only the results for models incldiung interactions are shown

[2] Dutch has two definite articles: 'de' and 'het'

* cointegration