Twitter as a potential source for official statistics in the Netherlands*

Piet J.H. Daas[1], Marko Roos[1], Mark van de Ven[2]
Joyce Neroni[3]
[1]Statistics Netherlands, Methodology sector, CBS-weg 11, 6412 EX, Heerlen, The Netherlands
[2]Erasmus University Rotterdam, Econometric Institute, Burg. Oudlaan 50, 3062 PA, Rotterdam, The Netherlands
[3]Utrecht University, Department of Psychology, Heidelberglaan 1, 3583 CS, Utrecht, The Netherlands

Abstract
An increasing number of people is active in social media. Here, people voluntarily share information, discuss topics of interest, and contact family and friends. Since the response to the questionnaires of Statistics Netherlands continuous to decline we investigated the potential usability of the information exchanged in social media as a data source for official statistics. Because Twitter is used by a large number of people in the Netherlands and the pubic messages can be relatively easily collected, we started to investigate the content of Twitter-messages. We collected the messages in various ways, classified the topics discussed and looked at the usability of the information from an official statistics point of view. User oriented message collection was found the best approach for our purposes. Identification of the topics discussed in the 12 million messages collected was done in two stages. First the topics in all the hashtag containing messages were determined and messages were classified. Next, a random sample of the non-hashtag containing messages was classified. The results revealed that a considerable amount of the messages collected, around 50%, could be of interest.

Key Words: Social networks, Data collection, Classification, Information exchange, Topic identification.

1. Introduction

Traditionally, sample surveys are used by National Statistical Institutes to collect data on persons, businesses, and all kinds of social and economical phenomena. During the last 30 years, more and more statistical institutes have gradually been replacing survey data with administrative data. This shift is predominantly caused by the wish to decrease the response burden on the data providers and the desire to produce statistics of sufficient quality in a cost efficient way (Bethlehem, 2010; Snijkers, 2009). Apart from administrative data sources there are, however, also other sources of secondary

* The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. The work described in this paper was presented at the 67th AAPOR 2012 conference, May 17-20, Orlando, Florida, USA.

information available in the world around us that could -potentially- provide data of interest for producers of statistics (Daas et al., 2011, Groves, 2011, Roos et al., 2009). Nowadays, more and more information is processed and stored by many of the ubiquitous electronic equipment surrounding us, such as mobile phones and other electronic devices. In addition, the ever increasing use of the internet causes more and more persons (and companies) to leave their digital footprint on the web (Dialogic, 2008). All these sources of information could potentially assist the production of statistics in a way similar to administrative data (Wallgren and Wallgren, 2007) and could even provide information describing new social and economical phenomena! (Daas et al., 2011; Nordbotten, 2010 and 2011; Hourcade et al., 2009). In this paper we explore the potential of a particular digital data source: data from the social medium Twitter. We focus on two main components, the collection of twitter messages and the classification of topics discussed.

1.1 Social Media

In the recent years the use of text based social media has vastly increased resulting in millions of people broadcasting their thoughts and opinions on a great variety of topics (Kaplan and Haenlein, 2010). Especially for social statistics studying the opinions, attitudes, and sentiments shared in social media could be interesting. Currently for official statistics, data available in many different sources, often surveys, are combined in the hope that this would fill the gap of information needed (Schmeets and Te Riele, 2010). Nowadays there are numerous sources on the internet that could be used to deduce similar facts. Examples of this kind of data sources are weblogs ('blogs'), news sites, and public chat rooms. These sources are, however, not always easy to find and not indexed well, because the data is spread over an ever-growing number of domain names.

Fortunately, some sources have seen a huge growth in popularity over the last few years and can be more easily investigated. These sources are the so-called micro-blogs (Java et al., 2007; Kaplan and Haenlien, 2011). Examples of such services are Google+, Twitter, and Tumblr (Sterne, 2010). Because Twitter is used by many people in the Netherlands (ComScore, 2011; Fisher, 2011) and much of its messages are publicly available, meaning that people that are not a member of the senders network are able to read it, make it a very attractive source of information (Laniado and Mika, 2010; Miller, 2011; Pear analytics, 2010). Also, adding to its opiniating effect, twitter messages of both well known and less well known people receive intense attention from the established media.

To study the usefulness of micro-blogging messages from an official statistics point of view, we focussed on public twitter messages. Before the results of our exploratory study are presented, we start by providing the reader the essential background information on the features specific to Twitter and its messages. This knowledge is needed to fully comprehend and understand the approach followed and the choices made in the research described in this paper.

1.2 Twitter features

Via Twitter users exchange information in short text messages, with a maximum of 140-characters (called ‘tweets’), by means of a central server located at “http://www.twitter.com”. A user that creates an account on Twitter automatically gets assigned a unique identifier and must provide a username, full name, and email-address. In addition, the user is requested to enter personal information like location and a short biography (a description). These are optional free text fields that are visible to the outside world unless the user chooses to hide their profile details. Apart from creating and sending messages, Twitter also enables users to subscribe to receiving messages from other users (‘follow’ a user). Every time a particular user sends a new twitter message to the server, all users that ‘follow’ that user, receive a notification of that message on their personal Twitter login page. These relationships are not reciprocal.

There are three ways by which a user can distribute a message on Twitter. A message can be send to the general public, to the followers of a user only, or -as a direct message- to one of its followers. The public availability of the first two types of messages is affected by the Twitter privacy settings of the user. Enabling the privacy option only allows his/her followers to read (and receive) the messages of a user. Users with the privacy option disabled have a public profile which makes there messages available to all people with internet access. Everybody can read their messages by, for example, visiting the users Twitter page at twitter.com/#/username. Direct messages, from one user to another, are always private. For our studies only publicly shared information of Twitter-users was collected.

Twitter messages also have specific characteristics. Each message has a date and time stamp. Additional features are: i) replying to a specific user (by including the ‘@username’), ii) use a hash sign (#) to ‘tag’ a word to highlight one or more keywords in the message, iii) ‘forward’ messages to followers, with the ‘retweet’ option, iv) adding a link to a web address, and v) add location information by including Global Positioning System (GPS) obtained coordinates or another source of location information as an attribute. Users that enable the optional location feature in their profile, assure that all their messages will include the location information from which their messages are send.

1.3 Scope of the study

The primary goal of the study described in this paper was the identification of the topics discussed on Twitter by Dutch users in the Netherlands. From this general approach it was assumed that the amount of messages relevant for official statistics and the area(s) of potential use could be deduced. To enable topic identification, first a considerable number of twitter text messages needed to be collected. The paper therefore starts by describing the ways by which twitter text messages were collected from inhabitants of the Netherlands. This resulted in a dataset that was used to identify the topics discussed on Twitter. The latter proved challenging. The paper ends with a discussion on the issues identified, the challenges remaining, and the potential use of Twitter in the context of official statistics.

2. Data collection and methods used

During our studies only publically available twitter data was collected. The data was securely stored on a server. When the data collection period ended the data was stored on an internal secure environment with access limited access to the authors and completely removed from the server. Because the database contained personal data, such as the username, processing of the data was done in accordance with the rules stated by the Dutch Data Protection Authority (DDPA, 2001).

2.1 How twitter data was collected

Twitter data can be collected via various Twitter ‘Application Programming Interfaces’ (API’s). After signing in with a username and password, twitter data can be obtained via one of three API’s: Streaming, Search, and REST (Twitter developer, 2012). Some features overlap between the API's but there are also considerable differences. Since Twitter API's are constantly being developed, the current situation may differ. Our primary demand was completeness of the set of messages collected. To get a good overview of the topics discussed on Twitter in the Netherlands, it was essential that no specific group of messages or users was missed. Because of the costs involved and the fact that our budget for this kind of research is limited, the Streaming approach was excluded. Preliminary studies with the Search API revealed that the results predominantly included messages from users with many followers. Messages from users with few followers, such as our test messages, were hardly ever included.

The REST approach was the way to go. Here, the user identifier is the point of entry. It not only allows the collection of the messages but also enables the extraction of data from users like followers (and their identifiers), profile and location information of users, and more. Unfortunately, the REST API has some limitations on bandwidth usage. However, by spreading the request to the Twitter server over multiple user accounts, we were able to collect all the information needed. Use of the REST API forced us to use the following sequence of events: i) collect as many user identifiers as possible, ii) identify the Dutch users within this population, and iii) subsequently collect messages from the Dutch users.

2.2 User name collection

The network of interconnected Twitter users forms a graph, were each user is a node. Users are linked by 'followers' and 'following' relationships. The first are the links between a user and its followers and the latter the relation between the user and the people he/she is following. Since some users are followed by many followers, we decided to start traversing through the network via the user-followers relationships. Dutch user identifiers were collected by a breath-first algorithmic approach; a data collection technique referred to as 'snowball' sampling by statisticians (Biernacki and Walldorf, 1981). Downside of this approach is that users can be missed, especially those that do not follow any other Dutch users. This could be solved by additionally adding the user-following relationships. However, because it already took a considerable amount of time to travel through the user-follower graph (close to 4 weeks) and because the number of unique Dutch users identifiers collected was quite close to the amount expected (see below), we decided not to additionally include the user-following relations findings.

2.3 Location name filtering

Identification of Dutch Twitter users was done by looking at the content of the location field in their profile. Users with the words 'Nederland', Netherlands, Holland, or the name of a Dutch province or municipality included in their location field were all initially considered Dutch. For users with an empty location field, the value 'unknown' was stored. Regular expression matching was used to compare strings. Although this approach returned quite good results, it did not suffice for all cities. For instance, a lot of clearly English and American city names matched positively with the Dutch village named 'Fort'. A considerable number of other foreign places, such as 'Amsterdam, Missouri, USA' and 'Bergen, Norway', also matched with Dutch city names. We solved this by creating an exclusion list of locations and regularly checking the results obtained. The exclusion list only contained locations that were certainly not Dutch and contained words like Belgium ('Belgi'), Germany, Deutschland, and some other obvious non-Dutch countries. In future studies the selection could be even more refined by including a language-sensitive analysis. Since the location information included in the user profile was used, geographic coordinates need not be considered here.

2.4 Text message collection

For all users identified as Dutch the 200 most recent his or hers twitter messages were collected. We choose this approach for several reasons. First, it prevented that the twitter messages send by very active users would dominate the topics discussed. Preliminary studies by one of the authors had already indicated that the number of tweets per user tends to follow a Zipfian distribution; plotted on a double-logarithmic scale it follow a straight line. Other studies additionally suggested that the messages from these kind of users are likely to be monothematic (Trump, 2010). The second advantage of this choice was that it considerably reduced the burden on the server; up to 200 messages could be collected by a single request.

3. Results

3.1 Dutch Twitter users

Studies from others performed around the time that we started our research suggested that the expected number of Dutch Twitter users should be somewhere between 150.000 and 320.000 (Cheng et al., 2009; Schoonderwoerd, 2011). This range was partly caused by differences in the definition of active users. Since it can be expected that a considerable number of users only create an account to obtain information, and not for sharing, the number of Dutch users could even be considerably higher.

Username data collection started by manually selecting a Dutch user with a great number of followers from the top five of most popular Dutch Twitter users. Our choice was a well-known and popular Dutch politician who was very active on Twitter during the Dutch elections in 2010 (Schoonderwoerd, 2011). At the time she was included in our database, she had exactly 79,798 followers. Next, all user identifiers, usernames, location information and other public profile information of her followers were collected. Subsequently, the information for all not already included followers of these followers were collected, etc. This process was repeated 8 times: the point at which the number of new Dutch user identifiers became nearly depleted. Table 1 provides an overview of the total number and new user identifiers at each stage. It also includes the number of requests needed to collect the data, indicating the burden this approach took on the Twitter server.

All in all, at the end the user database contained a total of 4,413,391 unique identifiers. Of these, 380,415 users -close to 9%- were positively identified as Dutch based on the information in their location field. Quite a large group of users, viz. 1,661,467 (38%), had no information in their location field resulting in the classification ‘unknown’. The remaining 2,371,509 users had a description in their location field that was not positively matched to the list of Dutch location names. These were therefore classified as ‘other’. Users in this group either lived outside the Netherlands or had a fantasy name in their location field; such as: ‘on Mars’, ‘close to the beach’ and ‘behind you’.

A total of 41% of the Dutch users had a reference to the country the Netherlands included in their location name. The city name that occurred most was Amsterdam (11%). A quarter of the location names contained the names of one of the five major cities in the Netherlands; i.e. Amsterdam, Rotterdam, The Hague, Utrecht and Eindhoven. Nearly half of the Dutch users had the optional description field filled in. In this field users often provide a short biography.

Table 1: Results of the collection of Dutch Twitter users.

Depth	*Total number of unique user ID's*[1] *collected*	*Total number of unique Dutch ID's (% of total ID's)*		*Number of new Dutch ID's (% of total Dutch ID's)*		*Total number of requests to server*
0	1	1	(100)	1	-	799
1	79,799	42,582	(53.4)	42,581	(100)	83,340
2	1,248,185	224,876	(18.0)	182,294	(81.1)	377,457
3	3,588,569	354,639	(9.9)	129,763	(36.6)	512,213
4	4,257,527	377,011	(8.9)	22,372	(5.93)	533,841
5	4,388,462	379,837	(8.7)	2,826	(0.74)	536,674
6	4,406,615	380,246	(8.6)	409	(0.11)	537,127
7	4,411,495	380,364	(8.6)	118	(0.03)	537,258
8	4,413,391	380,415	(8.6)	51	(0.01)	537,311

[1] ID: Identification number.

3.2 Messages collected

Capturing up to 200 messages of each of the 380,415 Dutch users identified resulted in a total of 12,093,065 twitter messages. Remarkably, for 39% of the users no messages were returned. This could be caused by the fact that those users had a private profile (indicating that no public tweets are available), they never created tweets, or had removed all tweets. These reasons could not be discerned. From the users for which messages were collected both the content of the message and additional meta-information was stored. The latter enabled the identification of the data/time and location associated with the message, whether a message was authentic, forwarded (a 'retweet') or a reply, and whether the message included user names, hashtags, and/or links to web pages. The vast majority of the messages covered 2009 and the first nine months of 2010. The oldest message obtained was sent on 20-10-2006. A general overview of the metadata characteristics of the messages collected is shown in Table 2.

3.3 Topic identification

To get an impression of the topics discussed, we first focussed on the twitter messages containing hashtags. By prefixing a word with a hash symbol (#), users add context to their twitter message. The hash tagged word essentially becomes a key word (Efron, 2010) which indicates the topic the message is about or refers to. Of the total number of messages collected 1,750,074 tweets (14.5%) contained a single hashtag and 12,378 messages contained two or more (0.1%). Because of their small number and potential disturbing effect, the latter group was ignored. As users are free to use and introduce hashtags, a considerable number of unique hashtags occurred; in total 16,439. The distribution of the number of messages per unique hashtag was highly skewed; it very much resembled a Zipfian distribution. The 300 most frequently used hashtags comprised a quarter of the total number of hashtag containing messages. The 1000 most frequently used hashtags represented a bit more than 35%.

By manually grouping the messages in which the 1000 most frequently used hashtags occurred, with the themes over which Statistics Netherlands publishes statistics (CBS, 2012) as a starting point, a start was made with the identification of the topics discussed on Twitter in the Netherlands. As a result of this initial classification the list of themes was adjusted somewhat as it was found that some themes did not or were hardly ever assigned and some themes were assigned much more. The resulting set of themes discerned is shown in the first column of Table 3. In column two a short description is

Table 2: Metadata characteristics of the Twitter messages collected for Dutch users

Types of messages	*Total number*	*Percentage of total (%)*	*Relative percentage (%)*
All	12,093,065	100	-
With hashtags	1,762,452	14.6	100
1 hashtag	1,750,074	14.5	99.3
2 or more hashtags	12,378	0.1	0.7
With username (@username)	4,821,669	39.9	-
With hyperlink	1,631,709	13.5	-
Original tweets	8,736,685	72.2	100
no username	1,392,438	11.5	15.9
no hashtag	1,473,329	12.2	16.9
no username and hashtag	241,855	2.0	2.8
Replies and retweets	3,356,380	27.8	100
total replies	3,022,310	25.0	90
total retweets	334,072	2.8	10

given. To these groups, the remainder of the single hashtag containing messages were additionally assigned. The relative contribution of the single hashtag containing messages eventually obtained for each group is shown in the third column of Table 3. This column reveals that in the single hashtag containing twitter messages the topics Media and Other most frequently occurred, followed by Sports and Spare time related tweets. The Other group was predominantly composed of messages in which the hashtagged word was sentiment related; such as #happy and #sad. In addition to the manual classification, we also applied automated text analysis techniques to classify the hashtag containing messages. Here, the software program LingPipe was used with an implementation of the DynamicLMClassifier (Alias-i 2008). The results obtained confirmed our earlier findings.

To get an idea of the topics discussed in the 10,330,613 non-hashtag containing messages, we started with the automated text classification method developed for the hashtag containing messages. We expected that this method would also assign the non-hashtag containing tweets to the predefined categories. However, in contrast to the findings for the hashtag containing messages, the results obtained for the non-hashtag group were very ambiguous; even after much effort. We therefore decided to manually classify a random sample of the non-hashtag containing messages. A sample of 1050 messages was drawn and the main topic discussed was assigned to the themes identified before. The relative contribution of the sampled messages to these topics is shown in the fourth column of Table 3. The findings of the sample not only indicated the relative contribution of the topics discussed but also revealed why the automated text classification method did not work for non-hashtag containing messages. Manual classification showed that the majority of the non-hashtag messages belonged to the Other group (51%). The great diversity of words included in this large group of messages must have negatively affected the automated classification process. We therefore discontinued our automated text classification efforts for these messages.

Table 3: Classification of hash and non-hashtag containing Twitter messages of Dutch users.

Theme	*Description*	*With single hashtag (%)*	*Without hashtags[2] (%)*	*Combined results (%)*
Economy	Referring to economy, income and enterprises	5	2	2
Education	School, teaching and training related	1	3	3
Environment	Nature, environment and other 'green' issues	0	1	1
Events	Non-sport and non-political happenings	4	1	1
Health	Health and welfare related	1	3	3
Holiday	Referring to on leave activities and travelling	1	2	2
ICT	Information and communication technology related	7	2	3
Living	References to a location, municipality or country	4	1	1
Media	Dutch TV and radio shows (non political)	20	5	7
Politics	Political debates, leaders, parties and government	7	2	3
Relations	Related to social and human interactions	4	1	1
Security	Security, crime and justice related	0	1	1
Spare time	Activities of people when not working (not sports)	9	10	10
Sports	Sports, clubs, and sports events	13	6	7
Transport	Referring to traffic, commuting and transport	2	3	3
Weather	Weather conditions, forecasts and warnings related	1	1	1
Work	Employment and job related	3	5	5
Other	Rest group	18	51	46

[2] Findings derived from a random sample.

In many of the Other non-hashtag containing messages unclear topics were discussed. These kind of messages are referred to as ‘pointless babble’ in some studies (Pear analytics, 2010). We will use the more neutral expression of ‘non-statistical interest’ here. Apart from these kinds of messages, the non-hashtag containing tweets in our sample were predominantly found to be related to the themes Spare time, Sport, Media and Work. Combining the findings for the hash and non-hashtag containing messages revealed that Other (46%), Spare time (10%), Media and Sports (both 7%) and Work (5%) related topics were most mentioned in our dataset. This, however, does not mean that topics discussed by a small percent of the messages are not of interest; a single percent still represents around 120 thousand messages in our dataset.

4. Discussion

The results described in this paper reveal that on twitter topics of potential interest for official statistics are discussed. Topics for which twitter messages could provide information from an official statistics point of view are those that are related to work and politics (Tjong et al., 2012). Spare time activities and events are also interesting options

(Schmeets and te Riele, 2010). Our personal experiences from looking into the content of twitter messages mentioning these topics supports the idea that quite some of these tweets could be used to provide opinions, attitudes, and sentiments towards these topics. Because of the vast amount of messages created on Twitter in the Netherlands (ComScore, 2011), this opens up possibilities to collect a considerable amount of information in a quick way without any perceived response burden. Problem is discriminating the informative from the non-informative messages. Because of the relative size of the Other group (see Table 3), many twitter messages discuss unclear topics and, hence, will very likely disturb the automatic identification of the relevant messages. Perhaps, pre-selecting tweets by the occurrence of topic specific words can be used to reduce this disturbing effect.

Studies by Bollen et al. (2011) suggest an additional use of tweets. They performed sentiment mining on all twitter messages collected during a certain period. Interestingly, the sentiments obtained were found to be related to stock market developments; suggesting a potential relation with economic developments, which is also interesting from an official statistics point of view. This finding indicates that, despite the fact that only a (selective) portion of the population uses Twitter, tweets could potentially be used as an indicator for developments in other areas of interest. The work of O'Conner et al. (2010) is another example of this approach. It also demonstrates that -for some applications- topic identification is not required. It will be interesting to attempt a similar approach in the Netherlands. Perhaps, not only opinions, attitudes, and sentiments towards the economy in general should be studied but perhaps also attitudes towards companies or specific branches of industry.

However, collecting twitter messages and analyzing their content is not the same as using this information for statistics. This is certainly not an easy hurdle to take. Based on our current experience, we expect that it will be difficult to relate the Twitter-based findings to the (opinion of the) Dutch population as a whole without using any additional source of information. This is caused by that fact that i) not every Dutch citizen is active on Twitter, ii) the activity on Twitter varies per user, iii) it is likely that not all users can be identified based on the information they (voluntarily) provide, and iv) the collection of tweets is rather selective. Perhaps studying the additional profile information provide by around 50% of the Dutch users provides insight on ways to solve some of these issues. Alternatively, a random sample of the Dutch population could be requested to provide their Twitter username.

Although it is clear that Twitter is a potential interesting source of information, still a considerable amount of work needs to be done to enable its actual use for official statistics. Future studies will therefore focus on the background characteristics of Dutch Twitter users, the improvement of the automatic classification of topics discussed on Twitter (and other social media), and on the mining of sentiment in the messages collected.

References

Alias-i (2008) LingPipe 4.1.0, available at: http://alias-i.com/lingpipe.

Bethlehem J. (2010) Statistics without surveys? About the past, present and future of data collection in the Netherlands. Paper for the 2010 International Methodology Symposium of Statistics Canada, October 26-29, Ottawa, Canada.

Biernacki P., Waldorf D. (1981) Snowball Sampling: Problems and Techniques of Chain Referral Sampling, *Sociological Methods Research* 10(2), 141-163.

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

CBS (2012) Themes overview, webpage available at: http://www.cbs.nl/en-gb/menu/themas/default.htm.

Cheng, A., Evans, M., Singh, H. (2009) Inside Twitter: An In-Depth Look Inside the Twitter World. Report of Sysomos, June, Toronto, Canada.

ComScore (2011) The Netherlands Ranks #1 Worldwide in Penetration for Twitter and Linkedin, available at: http://www.comscore.com/Press_Events/Press_Releases/2011/4/The_Netherlands_Ranks_number_one_Worldwide_in_Penetration_for_Twitter_and_LinkedIn.

Daas P., Roos M., de Blois C., Hoekstra R., ten Bosch O., Ma Y. (2011) New data sources for statistics: Experiences at Statistics Netherlands. Paper for the 2011 European New Technique and Technologies for Statistics conference, February 22-24, Brussels, Belgium.

DDPA (2001) Privacy Audit Framework under the new Dutch Data Protection Act (WBP). Working paper of the Co-operation Group Audit Strategy, Dutch Data Protection Authority, April, The Hague, The Netherlands.

Dialogic (2008) Go with the dataflow! Analysing the Internet as a data source. Report for the Ministry of Economic affairs, May 13, Utrecht, The Netherlands.

Efron, M. (2010) Hashtag retrieval in a microblogging environment. Paper for the 33rd international ACM SIGIR conference on Research and development in information retrieval, July 19-23, Geneva, Switzerland.

Fisher, E. (2011) European detail map of Flickr and Twitter locations, available at: http://www.flickr.com/photos/walkingsf/5912946760.

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

Hourcade, J-C., Saracco, R., Neuvo, Y., Wahlster, W., Posch, R. (2009) Future Internet 2020, Call for action by a high level visionary panel. Report of the European Commission Information Society and Media, Brussels, Belgium.

Java A., Song X., Finin T., Tseng, F. (2007) Why we twitter: understanding microblogging usage and communities, Paper for the ninth workshop on Web mining (WebKDD) and first social network analysis workshop (SNA-KDD), Aug 12, San Jose, USA.

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

Kaplan A.M., Haenlien, M. (2011) The early bird catches the news: Nine things you should know about micro-blogging, *Business Horizons* 54(2), 105–113.

Laniado, D., Mika, P. (2010) Making sense of Twitter, Paper for the 9th International Semantic Web Conference, November 7-11, Shanghai, China.

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

Nordbotten, S. (2010) The Use of Administrative Data in Official Statistics – Past, Present, and Future – With Special Reference to the Nordic Countries, in Carlson, Nyquist and Villani (Eds), *Official Statistics – Methodology and Applications in Honour of Daniel Thorburn*, Stockholm University, Stockholm, Sweden, pp. 205-223.

Nordbotten, S. (2011) Use of Electronically Observed Data in Official Statistics, Paper for the 58th Session of the International Statistical Institute, August 21-26, Dublin, Ireland.

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.

Pear analytics (2010) Twitter Study, August 2009, San Antonio, Texas, USA,.

Roos, M.R., Daas, P.J.H., Puts, M. (2009) Innovative data collection: new sources and opportunities (in Dutch). Discussion paper 09027, Statistics Netherlands, Heerlen.

Schoonderwoerd, N. (2011) Top 418.621 of Dutch Twitterrers (in Dutch), available at: http://nl.twirus.com/details/blog/731.

Snijkers, G. (2009) Getting Data for (Business) Statistics: What's new? What's next? Paper for the 2009 European New Technique and Technologies for Statistics conference, February 18-20, Brussels, Belgium.

Sterne, S. (2010) *Social Media Metrics: How to Measure and Optimize Your Marketing Investment*. John Wiley & Sons, Hoboken, USA.

Schmeets, H., Te Riele, S. (2010) A decline of social cohesion in the Netherlands? Participation and trust, 1997-2010. Paper for the International Conference on Social Cohesion and Development, January 20, Paris, France.

Tjong, E., Sang, K., Bos, J. (2012) Predicting the 2011 Dutch Senate Election Results with Twitter, Paper for the EACL 2012Workshop on Semantic Analysis in Social Networks, April 23, Avignon, France.

Trump, T. (2010) Types of Twitter users, Paper for the General Online Research conference 2010, May 26-28, Pforzheim, Germany.

Twitter developer (2012) Frequently Asked Questions, webpage available at: https://dev.twitter.com/docs/api-faq#how.

Wallgren A., Wallgren B. (2007) *Register-based Statistics: Administrative Data for Statistical Purposes*. Wiley, Chichester, UK.