Detecting the number of freelancers in the Netherlands using Twitter data

K.M. El Assal
University of Twente
s1097539
k.m.elassal@
student.utwente.nl

R. Kokkelkoren
University of Twente
s1617761
r.kokkelkoren@
student.utwente.nl

W.G. Oude Elferink
University of Twente
s1225588
w.g.oudeelferink@
student.utwente.nl

Keywords

Big Data, ZZP, Freelancer, Independent Contractor, Twitter

ABSTRACT

The Dutch Central Bureau of Statistics (CBS) keeps record of all kinds of statistics that are useful for the Netherlands. One of those statistics is the number of self-employed citizens without employees, also known as freelancers. The CBS wants to know if it is possible to use social media to determine the amount of freelancers in the Netherlands. Using MapReduce with semantic analysis of Twitter messages originating from a dataset of twiqs.nl, we try to determine the amount of freelancers active on Twitter. We then correlate that data with the data that is currently available at the website of the CBS. This paper describes how our semantic analysis was applied. We found a correlation of 0.864, which indicates that our semantic analysis was not sufficient to accurately determine the amount of freelancers in the Netherlands using Twitter but that it could be used to indicate a certain trend within the number of freelancers.

1. INTRODUCTION

This paper describes the research which will try to determine the number of freelancers in the Netherlands using Twitter data. The research is instigated by the course Managing Big Data from the University of Twente. This short paper will elaborate on the need of determining the number of freelancers in the Netherlands, and will also describe how this will be achieved and what the results were.

The Dutch Central Bureau of Statistics (CBS) keeps record of all kinds of statistics that are useful for the Netherlands. They report on many themes, such as health and well-being, international trade, population, and labor. One point of interest in the latter theme is the number of citizens that are self-employed without employees: freelancers ("ZZP'ers"). Every profit-driven entity can use social media as a marketing medium. Freelancers, specifically, need to reach many potential customers with relatively small effort and practically no cost. The CBS now wants to know if it is possible and feasible to use social media to identify and determine the number of freelancers in the Netherlands.

The focus of this research lies on the possibility of identifying freelancers based on their Twitter messages. The available Twitter data will be analyzed using the `Map/Reduce` method on a Hadoop cluster. This method will be used because it allows flexible development which is needed for this type of analysis. In addition, the researchers are more familiar with this type of method.

The analysis will be done using multiple methods which will result in a certain polarity of a Twitter message. The polarity of a message will be used to determine whether a user is a freelancer. The polarity will be calculated using methods such as semantic analysis. The semantic analysis will consist of determining whether the Twitter messages contain words that are affiliated with ZZP'ers. For example, if the Twitter message contains the words ZZP or KvK, it will receive a higher polarity. The results will contain only Twitter messages with a polarity higher than zero, indicating that one of the methods found an indication that the message was related to a freelancer.

The analysis will be performed on the Hadoop cluster made available by the Unverisity of Twente. The Twitter data will consist of the Twitter messages from 2011 till 2015 and was made available by `twiqs.nl` [10]. The Twitter data only contains messages originating from the Netherlands.

The output of this research will consist of a graph displaying the expected number of active ZZP'ers in the Netherlands. These graphs will be compared to those of the CBS which hold the true data for the period observed. A correlation between the graphs will show how well the Twitter analysis performs.

2. RELATED WORK

Detecting the number of zzp'ers (freelancers) using social media has not been attempted before by other researchers. There is however research done on detecting influenza epidemics [4], detecting communities with common interests [8], detecting air pollution [6] and on the demographics of twitter users [9]. These papers have in common that they all use a large number of tweets in order to detect certain statistics.

Aramaki et al. [4] make use of a twitter corpus to detect the number of people suffering from influenza. The benefits of using twitter data are the large number of messages including the word "influenza" and the fact that twitter enables

real time surveillance. To filter the twitter messages from "negative messages" where the word flu is not used in a sentence where the user suggests he has the flu, a support vector machine (SVM) is used. This classifier is trained using hand annotated positive and negative twitter messages. After filtering the number of remaining messages shows a correlation ratio of 0.89 with the real number of people suffering from influenza, outperforming the state-of-the-art method.

Jiang et al. [6] compare the air quality index (AQI) of Beijing with the number twitter messages about air pollution in Beijing. Where they first filter the messages for advertisements and indoor air pollution. They show that the number of messages show a correlation with the AQI. They then qualify each message as being negative or positive with respect to the air pollution. The frequency of positive and negative messages are used as features for a gradient tree boosting algorithm which improved the correlation with the AQI.

Lim and Datta [8] use Twitter to find communities with common interests. As a starting point, their approach uses Twitter accounts of celebrities that are a good example of a category of interest. Next, they analyze how the followers of those celebrities are networked together. Finally, they also analyze several characteristics of the found communities. According to Lim and Datta, there have been no other studies on the detection of communities with common interest on Twitter before theirs. Their method proved to be successful in several major categories and they observed how community structures become more connected and cohesive. Their approach was meant to be used as a tool for the implementation of target advertising.

One of the questions Mislove et al. try to answer is "who are the Twitter users?" [9]. They do that by analyzing geography, gender, and ethnicity. They concluded that the Twitter user base is not representative for the U.S. population: Twitter users are mostly located in densely populated regions, are male and the ethnicity distribution is highly skewed.

Another research which used data from social networking sites to examine a specific topic was published by Durahim et al. [5] In this paper Durahim et al. determine if it is possible to calculate a Gross National Happiness (GNH) for Turkey using data from Twitter. In order to validate their findings they compared the results against the Turkey Life Satisfaction Survey (LSS). The GNH was calculated using semantic analysis, this method assigns positive or negative polarity to certain words. The resulting polarity of the tweet message was determined by calculating the sum of all negative and positive polarity words. Using this method it was possible to determine the general positive or negative sentiment of the Turkeys population of different regions during certain time intervals. However it is still debatable how accurate this information is and if it can clearly represent the entire population.

Kashyap et al. published a similar paper as Durahim et al. in which they described a method to use Twitter data to monitor the population's health [7]. The basic analysis consisted of three parts. The first part consist of a semantic analysis on common words which are related to either good health practices or bad. The second part determines the sentiment of the user using the *Sentiment40* software. This is related to the fact that happier people are usually also healthier. The final part determines the positive or negative health polarity by analyzing the location in which the tweet was send. The research showed that monitoring the population's health is relatively effective.

3. MATERIALS AND METHODS

The method used is a text scanning algorithm which parses every tweet and searches for certain key words to determine the polarity of a tweet.

We did our research using Java runtime environment version 7 on the Hadoop cluster managed by the Centre for Telematics and Information Technology (CTIT) research group at the University of Twente. Our Java program uses the org.apache.hadoop library, version 2.6.0-cdh5.4.0, which equals the Hadoop version run on the cluster, and the org.json library, version 20151123.

The dataset consists of around 40% of all the generated dutch tweets between January 2011 and November 2015 in JSON format. This data is processed on the Hadoop cluster using the `Map/Reduce` method. As a first approach all tweets which contain the word zzp are found and are aggregated on a per day per unique user basis such that a user which has two tweets containing the word zzp in one day is only counted once. Furthermore the number of messages in the dataset are aggregated per day such that the data can be normalized. Next to this the data needs to be normalized for the number of twitter users, this data is collected from [3].

For the second approach we analyze every tweet on several key words, in the map phase every tweet is processed according to these rules:

- For each of the mentions "StZZPNederland" or "ZP-netwerk" in a tweet the polarity is increased by one.
- For each of the words "zzp","freelancer"," var ", or " var-" in a tweet the polarity is increased by one.
- For each of the hashtags "zzp" or "freelancer" in a tweet the polarity is increased by one.
- If the tweet contains the word "vacature" the tweet is ignored.

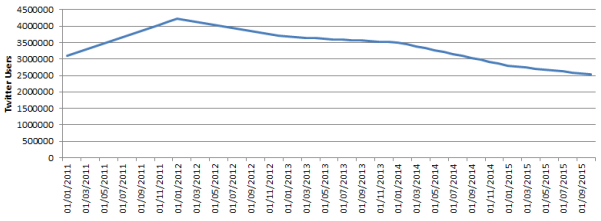

Figure 1: The number of twitter users.

If a tweet is considered to be a tweet from a zzp'er the polarity of the corresponding user is increased by one. If the polarity of a user in a month is higher than a certain threshold the user is considered to be a zzp'er and the count of zzp'ers is increased by one for that month.

The polarity threshold is different every month because the number of messages in the dataset (figure 2) and the number of twitter users (figure 1) is not the same for every month. The polarity threshold should be such that a user is counted as zzp'er if a certain percentage of his tweets in a

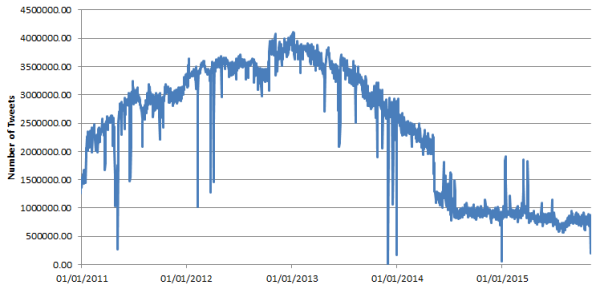

Figure 2: The number of tweets in the dataset per day according to [3].

month is a zzp tweet. To calculate this we need the number of tweets in the dataset per month and we need the number of twitter users. The average number of tweets per user per month can be found in figure 3.

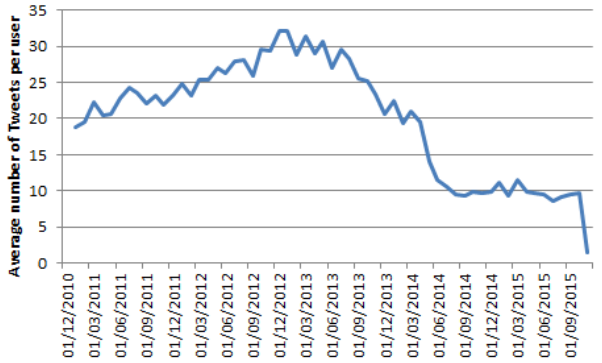

Figure 3: The average number of tweets per user per month.

From the figure it becomes clear that if we set the threshold to 40% then we need to find at least $0.4 * 30 = 12$ zzp related tweets per user in March 2013 while we need only $0.4 * 10 = 4$ tweets in April 2014.

4. RESULTS AND DISCUSSION

Since we need to know how accurate the results are, a comparison against the CBS data of the number zzp'ers should be made. In figure 4 the number of zzp'ers according to the CBS over the period 2008 till 2015 can be found. We can expect a gradual increase in the number of detected zzp'ers such that the number of zzp'ers in 2015 is 12.8% higher than in 2011.

Using the method of counting the occurrences of the number of tweets containing the word zzp per day and normalizing these occurrences by dividing them by the number of tweets in the dataset from that day and dividing this number to the number of twitter users will produce the graph as seen in figure 5. It can directly be found that the word zzp gets much more popular half way through 2014. This however means that when you compare the graph to that of the real number of zzp'ers in the Netherlands, there is no direct correlation visible within the graph. The cause for this is probably an increase in media attention which made the term zzp a more popular term on twitter. Furthermore an

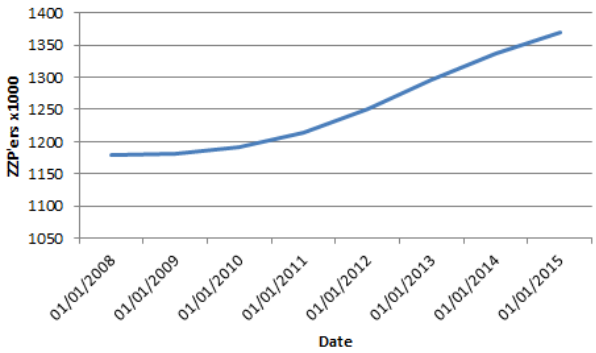

Figure 4: The number of zzp'ers in the Netherlands according to the CBS [1] [2]

increase in information and job offers for zzp'ers on twitter can be found in this same period as well. And finally by analyses of the tweets themselves it is found that zzp'ers do not actively use twitter to promote themselves.

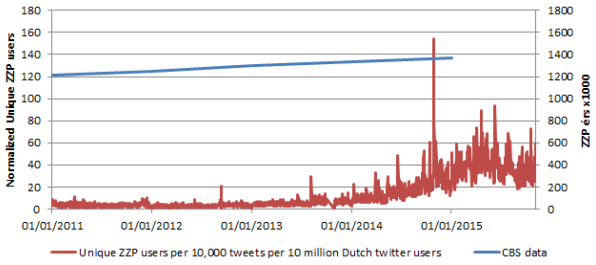

Figure 5: The number of zzp'ers according to the normalized count on twitter messages containing the word zzp per unique user per day per 10 million twitter users

The second approach defines a zzp'er as a user which has at least an x number of tweets with a high polarity that month. Where x is defined as 40% of the average number of tweets per month per user as seen in figure 3. Normalizing this by the number of twitter users in that month results in the graph as seen in figure 6. As can be seen directly, the same problem as in the first approach occurs. The popularity of the term zzp has too big of an influence.

Although at first glance there is no correlation between the CBS data and the research methods, the exact correlation still needs to be determined. The correlation between the sets of data is determined using the Pearson r coefficient which is calculated using equation 1. The Pearson r coefficient for both methods are shown in table 4.

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} \quad (1)$$

The Pearson r coefficient shows that there is a higher correlation between the different research methods and CBS data then visible within the graphs. It also shows that method 2 has a higher correlation with the CBS data then the first method. The high correlation between the normalized results method 2 and the CBS data shows that twitter analysis might be a valid method for identifying freelancers.

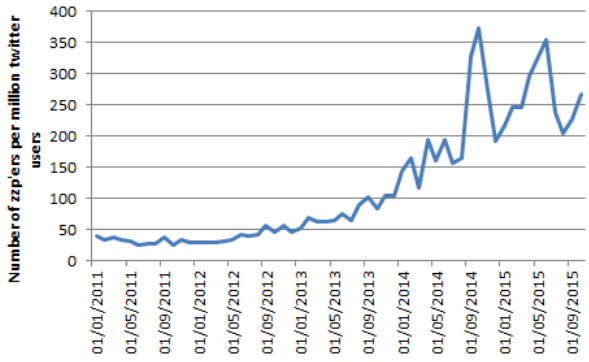

Figure 6: The number of zzp'ers per million twitter users per month according to the count on the number of users with a high polarity that month.

It is however doubtful that the trend continues to be highly correlated to the number of freelancers because it is unlikely that a real increase in zzp'ers of 12.8% will keep resulting in a factor 5 more tweets about freelancers as found in the results.

	Pearson Coefficient r
Method 1	0.59592239122772628
Method 2	0.86415139067500124

Table 1: Correlation coefficient of the research methods

5. CONCLUSION

This research tries to determine the number of freelancers within the Netherlands by performing analysis on Twitter data. Two different methods were used within this research, one which counted the number of Twitter users mentioning zzp each day. The second method performed thorough semantic analysis on each Twitter method and calculated a certain polarity of each user whether it is a freelancer or not. The second method had the highest correlation with the available data from CBS with a Pearson coefficient of 0.864.

This correlation coefficient indicates that thorough semantic analysis cannot be used to exactly determine the number of freelancers within the Netherlands. This is also related to the fact that twitter users are not a good representation of the Dutch population and due to the recent media attention for zzp'ers. However, the research method can be used to indicate a certain trend within the number of freelancers. It is therefore very likely that Twitter could be used as a valuable input source to get a current overall trend of the number of freelancers in the Netherlands.

Although the second research method already has a high correlation with the actual numbers of freelancers, there are still some aspects of the analysis which needs to be improved. Further research needs to be carried out before this method can be effectively used.

6. REFERENCES

[1] Centraal bureau voor de statistiek (2014). Arbeidsrekeningen; arbeidsvolume naar bedrijfstak en geslacht; 1969-2012.

[2] Centraal bureau voor de statistiek (2015). CBS: Meer mensen flexibel aan de slag.

[3] Cijfers aantal nederlandse gebruikers twitter in 2015. `http://www.buzzcapture.com/2015/06/cijfers-aantal-nederlandse-gebruikers-twitter-in-2015/`. Accessed: 2016-01-06.

[4] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics, 2011.

[5] A. O. Durahim and M. Coşkun. # iamhappybecause: Gross national happiness through twitter analysis and big data. *Technological Forecasting and Social Change*, 99:92–105, 2015.

[6] W. Jiang, Y. Wang, M.-H. Tsou, and X. Fu. Using social media to detect outdoor air pollution and monitor air quality index (aqi): A geo-targeted spatiotemporal analysis framework with sina weibo (chinese twitter). *PloS one*, 10(10):e0141185, 2015.

[7] R. Kashyap and A. Nahapetian. Tweet analysis for user health monitoring. In *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*, pages 348–351. IEEE, 2014.

[8] K. H. Lim and A. Datta. Finding twitter communities with common interests using following links of celebrities. In *Proceedings of the 3rd international workshop on Modeling social media*, pages 25–32. ACM, 2012.

[9] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Understanding the demographics of twitter users. *ICWSM*, 11:5th, 2011.

[10] E. Tjong Kim Sang and A. van den Bosch. Dealing with Big Data: the Case of Twitter. *Computational Linguistics in the Netherlands Journal*, 3:121–134, 2013. ISSN: 2211-4009.