Estimating the Actual Relocation of Dutch People Based on `Wish to Move' Messages on Twitter

Gao, H.

Estimating the Actual Relocation of Dutch People Based on 'Wish to Move' Messages on Twitter

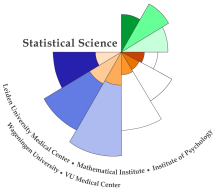
Han Gao (s1884352)

First advisor: Prof. Dr. Wessel Kraaij
Second advisor: Prof. Dr. Peter Grnwald
External advisor: Dr. Piet J.H. Daas

MASTER THESIS

Defended on October 9th, 2018

Specialization: Data Science



STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Abstract

Currently the 'wish to move' to another house of the Dutch people is measured through the WoON survey conducted every three years. A more frequent way of measuring is wished for to improve policy making in housing. Nowadays, people express their 'wish to move' on social media. In this research, it was found that certain features derived from tweet texts distinguish 'wish to move' tweets from others. The best logistic regression classifier developed in this research achieves an F1-score of 0.556 in identifying 'wish to move' tweets indicating that it is possible to timely keep track of the proportion of the 'wish to move' proportion of the Dutch population active on Twitter. Further, it is found that actual relocation can be identified by following 'wish to move' users. By engineering features through aggregating their subsequent tweets, classifiers were established to automatically determine if a 'wish to move' user relocated in the follow-up period. The best logistic regression classifier can determine if 'wish to move' users relocated in the two subsequent years with an F1-score of 0.701. With it, the proportion of 'wish to move' users who actually relocated later can be estimated.

Contents

1 Introduction

Traditionally, national statistical institutes, including Statistics Netherlands (CBS), use surveys as the main tool for the production of official statistics. However, the traditional survey is challenged by substantial cost, low response rate and potentially decreasing levels of funding [31]. With these concerns, statistical institutes try to explore alternatives. Nowadays, in the modern digital age when people increasingly share diverse and broad aspects of their daily lives on social media in real time [22], statistical institutes start to investigate social media data as an alternative. Especially in the Netherlands, 70% of the internet users aged above 12 are active on social media [4] and over 3 million public messages are produced everyday [8]. Inspired by this phenomenon, researchers at CBS have successfully conducted and are conducting projects using social media data to produce official statistics.

In this research, we perform a case study and try to investigate the pros and cons of social media analysis as an alternative to the traditional survey.

Knowing the 'wish to move' of Dutch residents is valuable for policy making in housing supply and the residential environment. Currently the 'wish to move' is measured every three years through the Netherlands' Housing Survey (WoON) on a random sample of the Dutch population consisting of more than 60,000 people [3]. CBS and the Ministry of Internal Affairs are interested to find an alternative method by which the frequency of measurement could be improved.

Recent research has shown that Dutch people produce, publicly available, social media messages in which they express their 'wish to move' to another house [10]. The objective of this research is to investigate whether social media data can be used to estimate Dutch people's relocation based on their 'wish to move' and whether this would enable a more frequent measurement than three-yearly.

However, a 'wish to move' is not always fulfilled as expected. Its realization is affected by several factors e.g. the strength of the wish, household, housing and regional characteristics [12]. Therefore, it's necessary to first investigate the link between the 'wish to move' and actual relocation on social media. Accordingly, the relocation of people who once expressed their 'wish to move' on social media is studied. The ultimate goal of this research is to determine if it is possible to develop a classifier, based on tweets of people

expressing ‘wish to move’ on social media and their background information, to determine their actual relocation.

In order to achieve the ultimate goal, the project is broken down into 3 parts. In the first part, classifiers were developed to identify ‘wish to move’ tweets. With the best classifier, ‘wish to move’ tweets produced during the period of the last WoON survey were identified and further checked manually. At the same time, background information such as age and gender of users of those real ‘wish to move’ were collected. In addition, tweets posted by those users were followed to see if a relocation actually happened in the subsequent two years. In order to distinguish if a tweet indicates relocation, the second part built classifiers in the same way as the first part. Further, the label ‘relocation’ or ‘non relocation’ for each user was manually checked so that it could be used as training output in the third part. The main use of the part is to obtain labels for the potential relocation tweets used to develop the final classifier. In the third part, based on all the information obtained in the previous two parts, i.e. ‘wish to move’ tweets, background information, follow-up tweets and relocation or not labels, classifiers were established to determine if a user expressing a ‘wish to move’ actually moved in the subsequent 2 years.

Since the goal of our research is to estimate Dutch people’s relocation, it is needed to check to what extent the underlying population on social media is representative of the Dutch population. Other challenges in our research include keeping the number of false positives and false negatives to a minimum when identifying ‘wish to move’ and actual relocation tweets.

2 Related Literature

Social Media Data and Official Statistics

Benefiting from its abundance and diversity, social media data is a good source to get an insight into public opinion and behaviour. Especially when we live at a time when the way we communicate is developing increasingly towards an online virtual world [28], social media stand out as a more and more appealing origin for the development of social indicators. On the other hand, concerning limitations of traditional survey, measures developed based on social media serve as supplements or even substitutes to the surveys [1, 32].

Recently, researchers have taken a step towards using social media data for the production of official statistics in various areas. By making use of Twitter messages, Gomide et al. proposed a four dimensions based surveillance methodology (volume, location, time and content) for detecting the outbreak of the tropical disease Dengue [14]. The method allows an accurate prediction of the level of dengue activity on a weekly basis. Antenucci et al. created indexes of labor market flows using tweets, showing the feasibility of measuring economic activity from social media postings [1]. Daas and Puts successfully developed a sentiment index from social media, which could be taken as an alternative for the Dutch Consumer Confidence Index measured by survey [9].

Studies also demonstrate the potential of social media data to be exploited in creating official statistical indicators in topics like healthcare performance, political preference, weather risk, etc. [6, 15, 30].

Event Detection

In the last years, numerous studies have been performed in the domain of event detection. The concept of an event, mentioned in the majority of studies, usually refers to a significant event, an event which may be discussed in the media [25] and is typically not of a personal nature. Besides, various terms are used in defining events from different perspectives. An open-domain future event is an upcoming popular event which could relate to any topic in any field [21]. A real-world event emphasizes a real-world occurrence mostly on a large or middle scale, while non event concerns messages about personal updates and random information[2]. There are many

more definitions regarding 'event' than those covered here since in practice researchers usually focus on custom-defined events which are especially interesting to them.

Hürriyetolu et al. invented a tool called Relevancer, able to automatically identify sets of tweets expressing a flexible sense of a key word in a tweet collection [17]. In his pipeline, unigrams and bigrams of any space-delimited sequence in tweets were used as features after text normalization. Then iterative clustering was applied to the tweet collection to pick coherent clusters for annotation. Subsequently, the labels acquired for the clusters enabled a Naive Bayes classifier trained for additional tweets classification. The performance of the tool on a case is 0.82 F1 score.

Kunneman constructed a system to detect future events and to identify their periodicity and emotions from a high-volume Twitter stream [19]. He established two approaches for event detection from two perspectives. The term-pivot approach, with a Winnow classifier based on a single feature i.e. the relative frequency by which an event is mentioned, can recognize significant events with 0.86 F1 score [20]. While the time reference-based approach, by pairing up future time expression and entity mentions co-occurring in tweets and ranking events by significance based on the date-term pair, is good at detecting open-domain future events with an F1-score of 0.548 (0.87 precision and 0.40 recall) [21].

In order to differentiate real-world events from non-events, Becker et al. first applied an incremental clustering to the messages in a Twitter stream and then, through designing a set of cluster-level features from temporal, social, topical and Twitter-centric perspectives, a support vector machine classifier gives an F1-score of 0.837 [2].

In contrast to detecting trending topics covered widely on social media, Walther and Kaisser presented an algorithm to discover real-world events in a given region in real-time which are often small-scale and localized (e.g. house fires or parties) as well as their location by monitoring Twitter streams [33]. First, tweets posted close in location and time are pre-selected and form into clusters. Subsequently, with four features selected representing textual and other aspects on cluster level, a pruned decision tree achieves an F1-score of 0.841 (0.832 precision and 0.850 recall) in determining whether clusters describe a real-world event or not.

Most researches in event detection are focused on significant events. There are few studies done on personal event detection from social media. Choudhury and Alani used multiple feature-based classifier to identify a number of personal events from Twitter [7]. An F1-score ranging from 0.754 to 0.862 was achieved on different events with the support vector machine classifiers constructed with linguistic and social interaction features. However, the classifiers they developed do not exclude tweets which resemble personal events but actually referred to an institution or other people instead of the tweet author himself.

In this research, the target events are 'wish to move' and actual relocation, both are personal. Unlike significant events which could be detected from mentions by multiple persons during specific periods, personal events usually happen independently without invoking messages of substantial volume close in time. Moreover, our focus is on self-reported 'wish to move' and relocation. Events referring to any institution or any other person should be excluded. This casts light on the challenge we face in the research.

3 Methodology

All social media data used is from the database of the Dutch company Coosto, which routinely collects public social media messages in Dutch from a whole range of social media platforms, e.g. Twitter, Facebook, Instagram and Youtube. Social media messages can be conveniently obtained by query via the web interface of Coosto. CBS has several licenses to use the data from Coosto.

Throughout the paper, all the experiments and analyses were done with data from Twitter. Although a pilot experiment was carried out including messages from Facebook [10], the work described in this research only focuses on Twitter data as the feasibility and reproducibility on Facebook in the Coosto database was negatively affected by the Facebook-Cambridge Analytica data scandal and subsequent changes in the privacy regulations of Facebook.

In order to determine the actual relocation based on 'wish to move' tweets, the discrepancy between the wish and the actual relocation is explored step by step, starting with identifying 'wish to move' messages.

3.1 'Wish to Move' Part

3.1.1 Data

The main data set used in the 'wish to move' part consists of 10,000 randomly sampled tweets containing either 'verhuis*' or 'verhuiz*' (both indicating move related words) as part of the text extracted through Coosto's database[1] [10]. These messages are produced on Twitter during 2014 to 2017. When extracting data from Coosto, those obviously originating from Belgian users with gps:Belgium or country:Belgium were excluded. This reduced the number of 'Dutch' tweets created by Flemish people.

Further, 969 tweets were randomly selected out of the main data set and annotated for training use. In order to obtain trustworthy labels, two annotators in parallel with a set of predefined rules determining real 'wish to move' message labelled the tweets either 'wish to move'(1) or non 'wish to

[1] Here '*' used as a wild card in query.

move'(0). Annotating resulted in 67 tweets labelled '1' and the rest labelled '0'. The kappa statistic was used to evaluate the interrater reliability[24] and it is 0.957 here.

3.1.2 Method

Existing Classifier

First, a classifier is needed to identify 'wish to move' tweets among those containing 'verhuis*' or 'verhuiz*'. The 'wish to move' tweet we target is a 'wish to move' message referring to a household rather than an institute/organization or someone else. Examples could be found in Table 1. For this step, a Logistic Regression (LR) classifier created by CBS with a 93% accuracy is available as the starting point. The classifier was trained with the 969 annotated tweets mentioned above. Word grams and other engineered features derived from the text were used. A word n-gram is defined as a contiguous sequence of n words from a given text[18]. A gram of size one is referred to as unigram and a gram of size two is called bigram, size three a trigram and so on.

Table 1: Examples of 'Wish to Move' and Non 'Wish to Move' Tweets

Tweet Example	**Classification**
'After living in Eindhoven for 1 year, I am now moving to Obdam near Alkmaar'	'wish to move' (referring to a person himself)
'No, I moved from my parents to my own house, I stayed in Hoorn'	non 'wish to move' (already moved)
'Symion is going to move https://t.co/j2PY6LebFK'	non 'wish to move' (referring to a company)
'Now in Overijssel Today #tvoost The Bathmen quad will move with parents and sisters to a larger house'	non 'wish to move' (referring to someone else)

Apart from the LR classifier, CBS also developed and tested various classifiers of machine learning techniques: Random Forest (RF), Support Vector Machine (SVM), Boosted Trees (BT) and Neural Network (NN). Although their testing accuracy is on the same level as LR, the LR classifier is preferred to them since it is interpretable. However, in spite of the 93% high accuracy of the LR classifier, it does not outperform the majority classifier

and it is useless as the minority class (wish to move) is the class of our interest. It suffers from heavily imbalanced training data (only 67 out of 969 i.e. 6.91% are positive). Therefore, the first task is to improve the classifier by including more positive 'wish to move' tweets in the training data and meanwhile to study other evaluation metrics suitable for unbalanced data.

Clustering

Hierarchical clustering together with manual annotation was used to find extra positive cases in the main data set (the one containing 10,000 tweets). Before clustering, some preprocessing was done. First, usernames and URLs in the text of those tweets were normalized into 'usrusr' and 'urlurl' respectively. Then, the texts were tokenized into 3-grams to 5-grams of character within word boundaries. Here, character n gram is a contiguous sequence of characters of size n. TF-IDF feature for each token was created for clustering similarity measure. After those steps, an agglomerative hierarchical clustering was performed on the processed tweets. The agglomerative hierarchical clustering is a clustering technique which starts with each data point, i.e. each tweet in our case, in its own cluster and then merge the two clusters closest in distance into a new one till the required number of clusters is reached [16]. Furthermore, the clusters were sorted in ascending order in term of the number of annotated 'wish to move' tweets contained for annotating. In this way, a positive 'wish to move' tweet could be identified efficiently with the least manual effort.

Model Training

Classifiers were then trained on the updated training data, i.e. the 969 annotated tweets plus extra positive tweets resulting from clustering.

Preprocessing. Duplicate Tweets were excluded from the updated training data. Preprocessing steps for modelling include removing non-ASCII sign (including emoji), punctuation and excessive white space. Stop words were kept and no stemming was performed, since for tweets, especially those short tweets, even stop words and original word forms are important in conveying information. Texts were tokenized into unigrams, bigrams and trigrams of word and TF-IDF of the grams were used as features. Sentiment label (neutral, positive or negative) for each message is automatically assigned by Coosto, which is used to indicate if a message expresses a neutral, positive or negative opinion. The sentiment of messages was used as a feature. Other

features than texts themselves were engineered including sentiment, length of text, number of words in text, company information, will move type sentence and non-move type sentence structure.The 'sentence structure' here refers to a sequence of certain contiguous words in a tweet. The details of all engineered features are listed in Table 2.

Table 2: Other Features in 'Wish to Move' Classifier

Feature	**Description**
sent	Opinion expressed in tweet 0:neutral, 1:positive, 2:negative
txtLen	Log10 transformed length of text with converting urls to 'http'
nonpersonal	Non-personal words ('company'/'firm'/'office'/'customer'/'shop') referring to an institute or organization rather than a household 0:no, 1:yes
numW1	Log10 transformed word count in the original text
numW2	Log10 transformed word count in the preprocessed text 'Word1/word2' form is split so that each single word could be counted in
persp	Count of person pronouns ('I'/'we'/'you')
willMove	Want to move type sentence ('move tomorrow'/'want(s) to move'/'going to move'/'move soon'/'finally move'/'new house' /'emigrate'/'move out') to identify sentence indicative for moving 0:no, 1:yes
nonMove	Non move type sentence ('new assessment'/'already done') to identify sentence indicative for non moving 0:no, 1:yes

Metric. To select an appropriate metric in our case, let us start with understanding the confusion matrix shown in Table 3.

Table 3: Confusion Matrix

		Predicted Class		
		0	1	
Actual Class	0	True Negative (TN)	False Positive (FP)	TN+FP=N
	1	False Negative (FN)	True Positive (TP)	FN+TP=P

The most commonly used classification metric, accuracy, is defined as the proportion of correctly classified cases. Accuracy concerns classification of both positive and negative classes. However, it can be misleading when

working with imbalanced data since most of the data could belong to one of the classes. In our case, the negative class (non 'wish to move' tweets) almost dominates the data but our interest is positive class ('wish to move' tweets). If we take accuracy as the performance metric, classifying all messages as non 'wish to move' already produces very good results (93% accuracy).

$$accuracy = \frac{TP + TN}{P + N}$$

Recall is defined as the proportion of positive cases which are classified correctly. It equals the number of true positives divided by the number of true positives plus the number of the false negatives. Recall represents the ability of a classifier to recognize real positives.

$$recall = \frac{TP}{TP + FN}$$

Precision is defined as the proportion of correctly classified positive cases relative to all the cases classified as positive. It equals the number of true positives divided by the number of true positives plus the number of the false positives. Precision represents how many positives classified by a classifier are real positives.

$$precision = \frac{TP}{TP + FP}$$

Specificity is defined as the proportion of negative cases which are classified correctly. It equals the number of true negatives divided by the number of true negatives plus the number of the false positives. Specificity represents the ability of a classifier to recognize real negatives.

$$specificity = \frac{TN}{TN + FP}$$

F1-score is defined as the harmonic mean of recall and precision. It considers recall and precision equally important by assigning them equal weights.

$$F_1 = 2 * \frac{recall * precision}{recall + precision}$$

Recall and precision do not take true negatives into account. *Balanced accuracy* is the mean of recall and specificity. It is an advanced accuracy metric which concerns both classes and deals with imbalanced data at the same time.

$$balanced\ accuracy = \frac{recall + specificity}{2}$$

The objective in this part is to recognize as many 'wish to move' tweets as possible. That is to say, we expect the classifier to identify true positives, i.e. correctly classified (labelled '1') 'wish to move' tweets, while controlling the number of false negatives, i.e. misclassified (labelled '0') 'wish to move' tweets and false positives, i.e. misclassified (labelled '1') non 'wish to move' tweets. To be specific, on the one hand, we expect a classifier which can identify as many real 'wish to move' tweets as possible so that those users with a 'wish to move' could be followed to see if any relocation happened. In that sense, false negatives should be decreased. Otherwise, it could lead to a set of real 'wish to move' users lost from the very beginning for follow-up. Missing 'wish to move' users at the first will lead to the loss in potential relocation users. Therefore, it is necessary to keep an eye on the metric recall so that we could keep unidentified 'wish to move' users to a minimum level. On the other hand, if the classifier gives false positives, i.e. takes non 'wish to move' as 'wish to move', then this will result in a set of users without 'wish to move' to be followed up unnecessarily. Therefore, it is important to take a close look at the metric precision to control false positives as much as possible. The above analysis illustrates that both precision and recall are important in evaluating the 'wish to move' classifier. Here, we take recall and precision as equally important in identifying 'wish to move' tweets. Therefore, the F1-score is taken as the main metric to select the best 'wish to move' classifier.

Other commonly used measures of accuracy in binary classification are Receiver Operating Characteristic (ROC) and Area Under Curve (AUC). The ROC curve represents the discriminative ability of a classifier at different threshold values. In our case, any tweet with predicted probability above a specific threshold value is classified as a 'wish to move' tweet. The ROC curve is produced by plotting recall against False Positive Rate (FPR), i.e.1-specificity, by varying the threshold value. As shown in Figure 1, the blue curve is an example of a ROC. The AUC is the area under the ROC curve. It is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

$$FPR = 1 - specificity = \frac{FP}{TN + FP}$$

However, AUC can become an inappropriate measure under certain situations. In the case of highly imbalanced data, AUC becomes insensitive[11]. More specific, in our case where negative class dominates the data, FPR hardly changes much since the total number of negatives is high. Second,

AUC concerns true negatives in its definition, which is not main concern in our case. Just as analysed before, our objective is to identify true positive cases while decreasing false positives and false negatives. Last but not least, AUC is a measure which summarizes the performance over various thresholds that researchers will rarely be interested in[23, 29]. Just like in our case, the threshold which could gives the highest accuracy is what we are looking for, while most of the possible thresholds do not make sense. Therefore, AUC is not selected to evaluate the classifier performance.

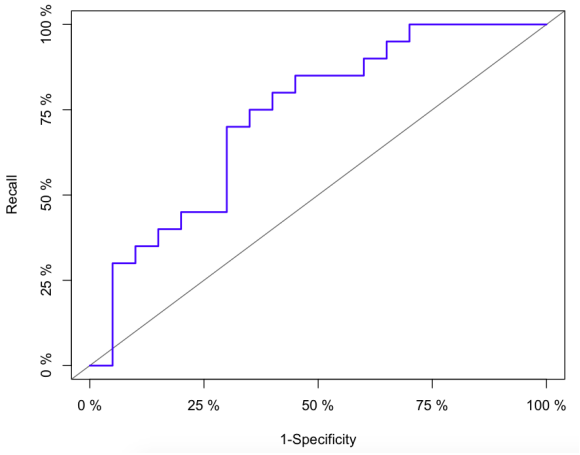


Figure 1: An Example of AUC

A stratified split with regard to the outcome variable 'wish to move' was made to the training data updated with more positive cases. Of the data 2/3 were used for training and 1/3 for testing. Then a stratified 10-fold cross validation was performed on the training part to tune hyperparameters. Different combinations of word grams were experimented, usually 1, 2 and 3. Additionally, considering the imbalanced data, weights for positive class ranging from 1 to 10 and various threshold values determining class label were experimented on training data as well. In the end, the performance of the selected classifier with specific hyperparameters, weight and threshold was evaluated on the 1/3 held-out data.

Besides, a classifier which classifies every tweet as 'wish to move' was used as the baseline. The reason that the majority classifier was not taken is that we concern positive (minority) class rather than negative (majority) class and the F1-score is the performance metric in this part. A majority classifier

which classifying each tweet as non 'wish to move' will result in 0 recall and not applicable precision, making it impossible to calculate F1-score.

Classifiers with different algorithms including LR, RF, SVM, BT and NN were constructed to find the best classifier. A brief introduction to these algorithms is given below.

Logistic Regression. LR takes a linear combination of dependent variables as input and estimates the probability of an outcome event using a logistic function.

Random Forest. RF is an ensemble method of decision trees. It is a modified bagging technique and works well especially for decision trees since they usually show high-variance and low-bias. The idea in RF is to build de-correlated trees by randomly selecting input variables to improve the variance reduction of bagging [13].

Support Vector Machine. In binary classification, SVM constructs a hyperplane which maximizes the distance to the nearest training data points of both classes. With kernel tricks, SVM is able to perform a non-linear classification by mapping data points into high-dimensional feature spaces.

Boosted Tree. BT is also an ensemble of weak learners, typically decision trees. However, unlike RF, it makes use of the boosting technique, which is fundamentally to reduce bias. Boosting produces a sequence of weak classifiers by applying the weak classification algorithm to repeatedly modified versions of the data. Then it combines those weak classifiers into a single stronger classifier.

Neural Network. NN encompasses a large class of models. The one we used in the research is the most widely used one called the single hidden layer back-propagation network.

3.1.3 Application to Tweets during WoON

The second step is to identify tweets produced during the period covered by the last WoON survey with the best 'wish to move' classifier. The survey was conducted from September 5th 2014 to May 30th 2015 and includes a question of 'wish to move' on it. Firstly all the tweets containing 'verhuis*' or

'verhuiz*' during that period were obtained from Coosto with those assigned as originating from Belgium excluded. Since the weekly proportion of 'wish to move' people is available from the WoON survey, we calculated the weekly proportion of 'wish to move' tweets identified by the classifier during the WoON survey to obtain an estimate for comparison.

Another objective of this step is to provide clean 'wish to move' users for the purpose of training in the last part. Hence, location information (if available) of all those 'wish to move' users in their profile were collected. Given their location information, only users living within the Netherlands were kept to further exclude Flemish people. Subsequently, the 'wish to move' tweets of the remaining users were manually checked by two annotators in parallel. In order to assist manual checking, hierarchical clustering was applied so that similar tweets in a cluster could be presented to annotators for checking. The preprocessing steps of the clustering is same as the one in the first part. After that, we got a set of users who were manually confirmed as real 'wish to move' users by both annotators. Then 1,000 users were randomly selected out of those users to be used as the sample in the last part.

Besides, all the publicly available tweets posted afterwards by the 1,000 users were collected and their age and gender were estimated [27].

3.2 Actual Relocation Part

In the previous part, the 'wish to move' classifier was developed and with it those real 'wish to move' tweets produced during the WoON survey period were identified. This resulted in a set of (1,000) 'wish to move' users which could be followed up from the time point of their 'wish to move' on to see if any relocation happened. For each user, the label relocation or not needs to be determined in this part so that it could be used as the output for training the final classifier in the last part.

In order to get those labels, the first step was to create a classifier that could indicate if a follow-up tweet containing 'verhuisd (moved)' or 'nieuw huis (new house)' refers to the user's relocation. Then, in order to ensure those labels clean to be taken as output labels for the purpose of training in the third part, manual checking was further applied.

In a word, the main use of the relocation classifier is to support the devel-

opment of the final classifier in the third part.

3.2.1 Data

Just as in the 'wish to move' part, a classifier is needed to identify real relocation tweets. The main data set used for model training in this part consists of 1,000 randomly sampled tweets containing either 'verhuisd' or 'nieuw huis'. They were extracted through Coosto's database with those clearly originating from Belgium excluded. These tweets were produced on Twitter during 2014 to 2017. All the tweets were annotated '1' for actual relocation and '0' for non-relocation by two annotators in parallel. The kappa statistic was used to test the interrater reliability.

3.2.2 Method

Table 4: Other Features in Relocation Classifier

Feature	Description
sent	Opinion expressed in tweet 0:neutral, 1:positive, 2:negative
txtLen	Log10 transformed length of text with converting urls to 'http'
nonpersonal	Non-personal words ('company'/'firm'/'office'/'customer'/'shop') referring to an institute or organization rather than a household 0:no, 1:yes
numW1	Log10 transformed word count in the original text
numW2	Log10 transformed word count in the preprocessed text 'Word1/word2' form is split so that each single word could be counted in
entity	Count of named entities(sequences of more than one capitalized word)
persp	Count of person pronouns ('I'/'we'/'you')
Moved	Moved type sentence ('have moved'/'has moved'/'just moved' /'now moved'/'completely moved'/'moved to a new house') to identify sentence indicative for moved 0:no, 1:yes
nonMoved	Non moved type sentence ('not moved'/'buy a new house' /'bought a new house'/'looking for a new house'/'found a new house') to identify sentence indicative for non moved 0:no, 1:yes

Preprocessing. The same preprocessing was applied to the 1,000 tweets as in the 'wish to move' part. Tweets were tokenized and unigrams, bigrams

and trigrams of word were used as features. Other features than word grams were created, including sentiment, length of text, number of words in text, company information, named entity, moved and non-moved type sentence. Details of all the engineered features are listed in Table 4.

Model Training. Just as in the 'wish to move' classifier, a stratified split in regard to the outcome variable relocation was done with the training data. Of the data 2/3 were used for training and 1/3 for testing. Then a stratified 10-fold cross validation was performed on the training part to tune hyperparameters. Different combinations of word grams were experimented. Additionally, considering the imbalanced data, weights for positive class ranging from 1 to 10 and various threshold values determining class label were experimented on training data as well. In the end, the performance of the selected classifier with specific hyperparameters, weight and threshold were evaluated on the 1/3 held-out data. A baseline classifier which classifies every tweet as a non-relocation was used.

Metric. In this part, the objective is to develop a classifier which can determine if a tweet containing 'verhuisd' or 'nieuw huis' indicates a relocation. It is worth pointing out that, unlike in the first part only focusing on 'wish to move' but not non 'wish to move' since the goal is to follow users wishing to move, here in this part both relocation and non-relocation are important. Both information determined from the follow-up tweets will be taken as output to train the final classifier in the third part. Therefore a metric considering both of the two classes should be chosen. The balanced accuracy was the one taken here to select the best relocation classifier.

Classifiers with different algorithms including LR, RF, SVM, BT and NN were constructed.

3.2.3 Application to WoON

All users expressing a 'wish to move' during the WoON period were followed to see if any relocation took place. Each user was followed for two years since his 'wish to move'. All their publicly available tweets containing 'verhuisd' or 'nieuw huis' during the period were selected and classified by the best relocation classifier.

In order to provide clean relocation labels for the purpose of training in the

last part, the labels assigned by the classifier were manually checked and improved. To assist manual checking, hierarchical clustering was applied so that similar tweets could be presented for checking. The preprocessing steps of the clustering is the same as the one in the first part. After that, for each of the 1,000 'wish to move' users, relocation information during the subsequent two years was obtained. This will be used as output labels for training in the last part.

3.3 From 'Wish to Move' to Actual Relocation

In the third part, by following tweets posted during the subsequent two years, a final classifier was established to determine if a user expressing a 'wish to move' on Twitter actually relocated.

3.3.1 Data

All the data used in this part were obtained from the previous steps. For each of the 1,000 'wish to move' users, there are three inputs and one output used for training the final classifier shown in Figure 2.

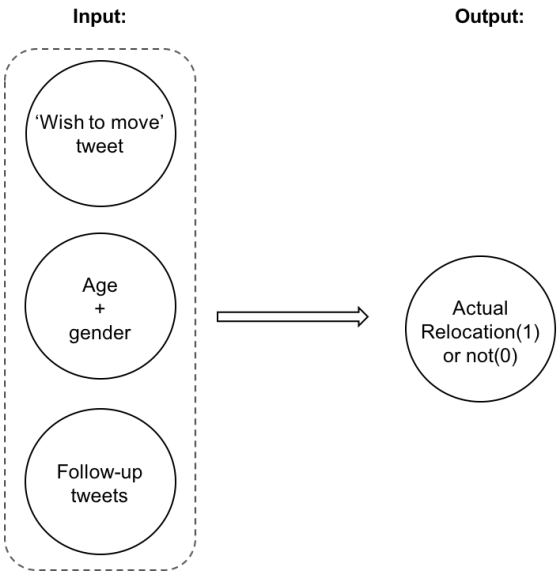


Figure 2: Input and Output of the Final Classifier

The first input are the 1,000 'wish to move' tweets during the WoON survey obtained in the first part. They were initially identified by the best 'wish to move' classifier and then confirmed manually to ensure they were correctly labelled as 'wish to move' messages.

The second piece of input is background information, i.e. age and gender of the 1,000 'wish to move' users.

The third input is their follow-up tweets in the subsequent two years. In order to get rid of abundant irrelevant tweets, only tweets related to relocation, i.e. tweets containing 'verhui*' or 'nieuw huis' were selected.

In the second part, for each of the 1,000 users, the information relocation or not was obtained by tracking his/her tweets during the follow-up period. Resulted actual relocation (1) or not (0) for each user is output used here for training. All mentioned information was taken to train the final classifier.

3.3.2 Method

Preprocessing

The same preprocessing as in the 'wish to move' part was applied to the input tweets. Tweets were tokenized and unigrams, bigrams and trigrams of word were used as features. There is a distinction in the input between the final classifier and the other two classifiers. The final classifier takes more than one tweet as input while the other two only take one tweet to make the prediction. Therefore, it naturally points to the fact that some aggregated feature could be introduced. Features created here other than word grams, age and gender include sentiment, length of text, number of words in text, company information, named entity, moved and non moved type expression. The details of them are listed in Table 5.

Model Training

Similar as in the relocation classifier, a stratified split with regard to the outcome variable relocation was done with the training data. Of the data 2/3 were used for training and 1/3 for testing. Then a stratified 10-fold cross validation was performed on the training part to tune hyperparameters. Different combinations of word grams were experimented. Additionally, considering the imbalanced data, weights for positive class ranging from 1 to 10

Table 5: Other Features in Final Classifier

Feature	**Description**
sentN	Normalized number of negative tweets
sentP	Normalized number of positive tweets
sentNeu	Normalized number of neutral tweets
txtLen	Normalized length of text with converting urls to 'http'
nonpersonal	Normalized number of non-personal words ('company'/'firm' /'office'/'customer'/'shop') referring to an institute or organization rather than a household
numW1	Normalized word count in the original text
numW2	Normalized word count in the preprocessed text 'Word1/word2' form is split so that each single word could be counted in
entity	Normalized count of named entities
persp	Normalized count of person pronouns ('I'/'we'/'you')
Moved	Moved type sentence ('have moved'/'has moved'/'just moved' /'now moved'/'completely moved'/'moved to a new house') to identify sentence indicative for moved 0:no, 1:yes
nonMoved	Non moved type sentence ('not moved'/'buy a new house' /'bought a new house'/'looking for a new house'/'found a new house') to identify sentence indicative for non moved 0:no, 1:yes
age	Age of user
gender	Gender of user 0:female, 1:male

and various threshold values determining class label were experimented on training data as well. In the end, the performance of the selected classifier with specific hyperparameters, weight and threshold were evaluated on the 1/3 held-out data. Besides, a classifier which classifies every user as relocated was used as the baseline here. The reason that the majority classifier was not taken is same as in the first part.

Metric. In this part, the objective is to develop a classifier which can determine if a 'wish to move' user will actually relocate. Similar as in the 'wish to move' part, the focus is on recognizing the positive class. Therefore, the F1-score is taken as the main metric to select the best final classifier.

Classifiers with different algorithms including LR, RF, SVM, BT and NN were also constructed. In addition, in order to see how varying follow-up periods affects the accuracy of identifying relocation users, different classifiers were developed and compared.

4 Results

4.1 'Wish to Move' Part

Clustering

With an extra 31 positive cases added through clustering and annotating, 98 out of 1,000 tweets in the updated CBS training data set were positive, i.e. the positive ratio increased from 6.91% to 9.80%.

Model Training

Classifiers of different combinations of features were developed. Their performance on the test data are shown in Table 6. It demonstrates that the classifier with unigrams, bigrams, trigrams and other engineered features achieves the best performance of an F1-score of 0.556.

Table 6: Performance of LR 'Wish to Move' Classifiers on the Test Data.

Model	F1	Recall	Precision
Existing Classifier	0.267	0.182	0.5
baseline	0.131	1	0.07
1 grams+other	0.535	0.594	0.487
1+2 grams+other	0.533	0.625	0.465
1+3 grams+other	0.551	0.594	0.514
1-3 grams+other	0.556	0.625	0.5

With the best classifier, 62.5% of 'wish to move' tweets can be recognized and 50% of tweets classified as 'wish to move' are real 'wish to move'. It is worthy to point out that compared with the existing LR classifier with $F1 = 0.267$, adding more positive cases in the training data did help to improve the performance, especially the recall increased from 0.182 to 0.625. In other words, adding extra positive cases to training data makes the classifier better in recognizing more 'wish to move' tweets. In addition, all classifiers outperform the baseline classifier. Adding word bigrams and trigrams to the classifier only taking unigrams and other engineered features improves its performance in recall and precision, respectively. Hence, the classifier constructed with unigrams, bigrams, trigrams and other features shows the best performance in the F1.

Classifiers constructed with other algorithms were trained on 1-3 word grams

and other features. Their performance are listed in Table 7. It shows that the RF classifier achieves an F1-score of 0.540 which is just slightly lower than the best LR classifier while other algorithms reach no more than 0.5 F1-score.

Table 7: Performance of Other 'Wish to Move' Classifiers on the Test Data.

Model	**F1**	**Recall**	**Precision**
RF	0.540	0.531	0.548
SVM	0.446	0.906	0.296
BT	0.5	0.563	0.45
NN	0.48	0.563	0.419

Table 8: Top 10 Most Important Features and Their Coefficients in the 'Wish to Move' Classifier

	Coefficient	
Feature	**Positive**	**Negative**
txtLen		-1.103
numW2		-0.669
willMove	0.551	
persp	0.390	
'tco'(part of URL)		-0.329
'verhuizen(moving)'	0.326	
sent	0.205	
'verhuist(moved)'		-0.179
'verhuisd(moved)'		-0.174
'beter(better)'		-0.146

Based on this result, we continued with the best LR classifier. The top 10 most important features and their coefficients in the best 'wish to move' classifier are listed in Table 8. It can be seen that a 'wish to move' tweet is usually shorter in length and contains fewer words than a non 'wish to move' tweet. Want to move type sentence (willMove) does help to capture 'wish to move' tweets well. First and second person pronouns (persp) and non-neutral sentiment (sent) are more associated with 'wish to move' tweets than non wish ones. Furthermore, terms like 'verhuizen (moving)' more likely indicate a 'wish to move', while 'verhuist (moved)', 'verhuisd (moved)' and 'beter (better)' more likely indicate a non 'wish to move'. A URL in a tweet is usually used for information sharing and it could be less likely personal event related, therefore it might appear more in non 'wish to move' tweets.

Application to Tweets during WoON

With the best 'wish to move' classifier, 21,142 tweets out of the total 179,758 ones containing 'verhuis*' or 'verhuiz*' during the last WoON survey were classified as positive. Weekly proportions of 'wish to move' based on the WoON survey and the 'wish to move' classifier are plotted respectively in the Figure 3. The blue line (left) is the weekly proportion in hundredth of Dutch people who certainly or probably wants to move based on the WoON survey. The red line (right) is the weekly proportion in millionth of 'wish to move' tweets identified by the classifier to all the tweets during the same period. There is a big difference between the order of magnitude of the two proportions and there is no certain similar pattern between the two proportions concluded from the figure. The correlation of them is -0.372.

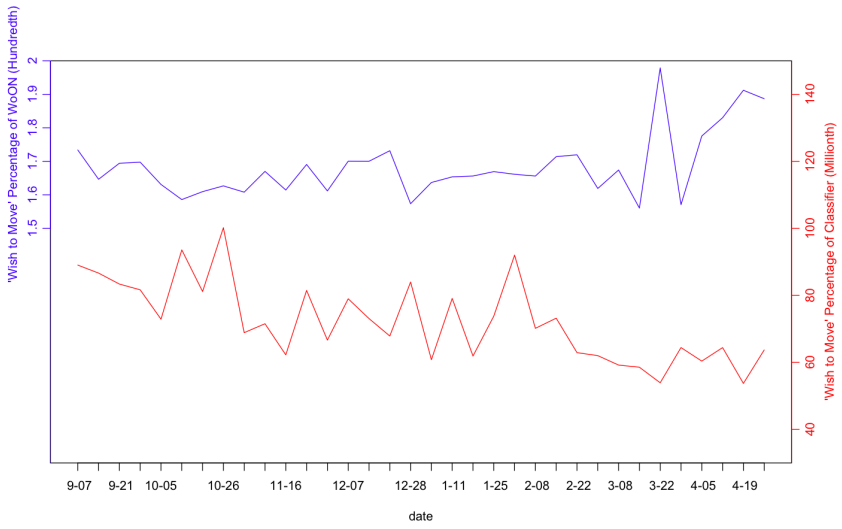


Figure 3: Weekly Proportion of 'Wish to Move' per WoON and per Classifier

Identifying 'Wish to Move' Users

There are 17,901 tweets after duplicates were excluded from the complete set of 'wish to move' (21,142) tweets. Some of these are still retweets. In that case, the first author of the tweet was identified as the user to follow up, instead of the user who later on retweeted. As shown in the faked example in Figure 4, it is the user 'userA' who first posted his wish to move and then the user 'userB' retweets it. Therefore, it is 'userA' that is to be followed.

@userB has retweeted:

RT @userA: I am going to move to Leiden ☺

Figure 4: The First Author in A Faked Retweet

Identifying the first author resulted in 12,756 unique users. Further given their location information in their profile, 10,098 users within the Netherlands were kept and 4,263 of them were manually confirmed as 'wish to move' users by two annotators. Finally, 1,000 users were randomly selected as the sample for training in the last part. All the analysis steps of identifying the 'wish to move' users during the WoON are illustrated in Figure 5.

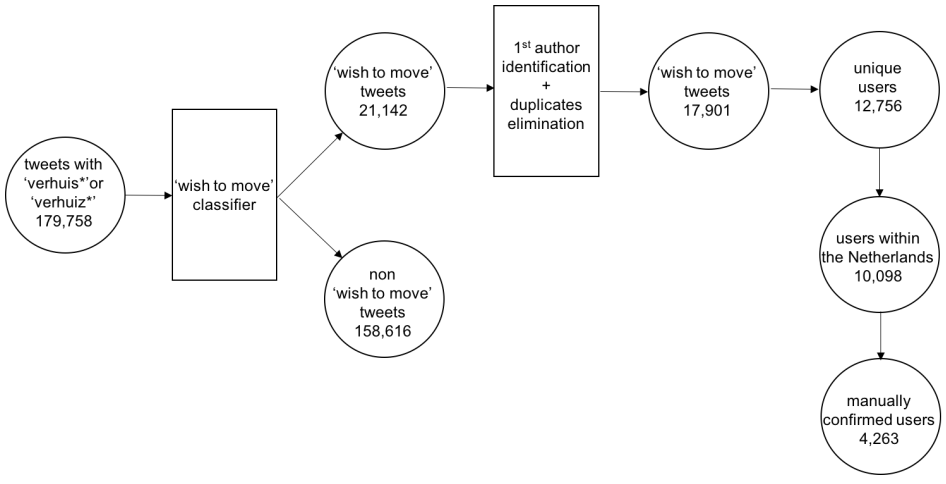


Figure 5: Identification Process of 'Wish to Move' Users During the WoON Survey Period

From the time point of their 'wish to move' tweets, the 1,000 users were followed up in the subsequent two years, i.e. all the relocation related tweets of each user after his 'wish to move' were collected for use in the subsequent steps. In addition, age and gender were estimated for those users.

4.2 Actual Relocation Part

The kappa statistic between the two annotators for the training data annotating is 0.722, which indicates that a substantial agreement was reached on what content can be annotated as a real relocation tweet. There are 162 out of the 1,000 tweets marked as real relocation where both of annotators agreed on the label.

LR classifiers were developed on different combinations of features respectively. As discussed above, balanced accuracy is used as the main performance metric to select the best relocation classifier. Their performance on the test data are shown in Table 9. F1-score, recall and precision are also listed for information. It shows that all the classifiers outperform the baseline classifier which classifies every tweet as a non-relocation. Adding word bigrams and trigrams step by step to the classifier only taking unigrams and other features does help improve the classifier in balanced accuracy. The classifier constructed with unigrams, bigrams, trigrams and other engineered features gives the best performance of a balanced accuracy of 0.792 and an F1-score of 0.560.

Table 9: Performance of LR Relocation Classifiers on the Test Data.

Model	**Balanced Accuracy**	F1	Recall	Precision
baseline	0.5	n/a	0	n/a
1 grams+other	0.715	0.496	0.574	0.437
1 and 2 grams+other	0.760	0.550	0.667	0.468
1 and 3 grams+other	0.775	0.544	0.740	0.430
1-3 grams+other	0.792	0.560	0.778	0.438

Classifiers constructed with other algorithms were trained on 1-3 word grams and other features as well. Their performance is listed in Table 10. It shows that the performance of all the algorithms are lower than the best LR classifier.

Table 10: Performance of Other Relocation Classifiers on the Test Data.

Model	**Balanced Accuracy**	F1	Recall	Precision
RF	0.744	0.569	0.574	0.564
SVM	0.748	0.566	0.593	0.542
BT	0.743	0.541	0.611	0.485
NN	0.698	0.483	0.519	0.452

Based on this result, we continued with the best LR classifier. The top 10 most important features and their coefficients in the best relocation classifier are listed in Table 11. It can be seen that non-moved type sentence structure does help capture non relocation information. Person pronouns can distinguish relocation tweets from non relocation ones: third person pronoun 'ze (she/they)' appears more in non relocation tweets while first and second person pronouns (persp) appear more in relocation tweets. Furthermore, terms like 'nieuw huis (new house)' and 'vorige (previous)' more likely indicate a relocation while 'leuk (nice)', 'goed (good)', 'zijn verhuisd naar (have moved to)' and 'is verhuisd naar (has moved to)' more likely indicate a non-relocation. A URL in a tweet is usually used for information sharing and it could be less likely a personal event hence it appears more in non-relocation tweets. Besides, the feature entity has its coefficient -0.609 in the classifier, which indicates that relocation tweets likely include more named entities.

Table 11: Top 10 Most Important Features and Their Coefficients in the Relocation Classifier

	Coefficient	
Feature	**Positive**	**Negative**
notMoved		-1.918
'leuk(nice)'		-1.656
'nieuw huis(new house)'	1.386	
'ze(she/they)'		-1.342
'vorige(previous)'	1.323	
'goed(good)'		-1.245
'zijn verhuisd naar(have moved to)'		-1.244
'is verhuisd naar(has moved to)'		-1.119
persp	1.068	
'tco'(part of URL)		-1.056

Application to WoON

During the 2-year follow-up, 405 out of 1,000 'wish to move' users posted 1,826 relocation related tweets (tweets containing 'verhuisd' or 'nieuw huis'). With the best relocation classifier, there are 241 users classified as actual relocated users. At the same time, by manually checking those follow-up tweets, there are 213 users identified as relocated users.

In addition, a short study of the location information provided by the users, whether on their profile or as metadata in their tweets, revealed that this approach should not be followed. Checking the content of tweets worked much better as, apparently, users do not update their location regularly or move outside the region indicated.

4.3 From 'Wish to Move' to Actual Relocation

In this part, 1,000 randomly selected 'wish to move' information was used to train the final classifier. The information we got for each of the 1,000 'wish to move' users from the steps outlined above includes the 'wish to move' tweet, background information (age and gender) and other tweets containing 'verhui*' or 'nieuw huis' posted in the subsequent two years. These three types of information serves as the input for training the final classifier. For each user, we also got the information whether he/she relocated by following his/her tweets in the second part. This is used as the output for training in this part.

LR classifiers were developed on different combinations of features respectively. The performance of these classifiers on the test data are shown in Table 12. It shows that all the classifiers substantially outperform the baseline classifier which classifies every user as a relocation. Adding word bigrams to the classifier only taking unigrams and other features does improve its performance in the F1-score. However, adding trigrams does not result any further improvement. The classifier constructed with unigrams, bigrams and other engineered features achieves the best performance of an F1-score of 0.715. Hence, we continued with unigrams, bigrams and other features.

Table 12: Performance of LR Final Classifiers on the Test Data.

Model	F1	Recall	Precision
baseline	0.356	1	0.216
1 grams+other	0.671	0.775	0.591
1 and 2 grams+other	0.715	0.831	0.628
1 and 3 grams+other	0.692	0.775	0.625
1-3 grams+other	0.701	0.775	0.640

By varying the follow-up period from 1 month to 1.5 years, different LR classifiers were developed. As shown in Figure 6, with increasing follow-up period, both the F1-score and the precision of the classifier increase. A longer period also results in higher recall except for 1-year follow-up showing a lower recall than 3-month and 6-month. Hence, we continued with the 2-year follow-up period.

Table 13: Performance of LR Final Classifiers with Different Period on the Test Data.

Model	F1	Recall	Precision
2 years	0.715	0.831	0.628
1.5 years	0.671	0.731	0.620
1 year	0.581	0.567	0.596
6 months	0.569	0.633	0.512
3 months	0.547	0.619	0.491
1 month	0.429	0.462	0.400

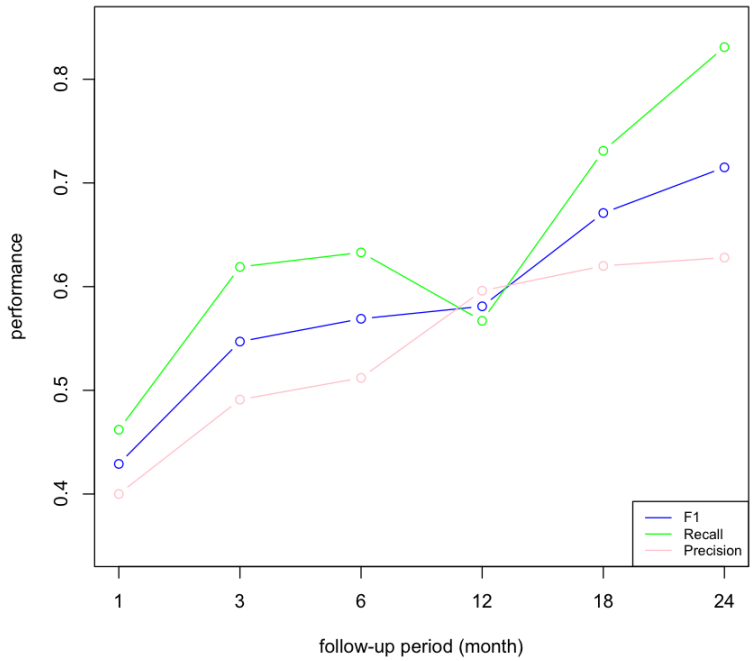


Figure 6: Performance of LR Final Classifiers with Different Period

Further, classifiers constructed with other algorithms were trained on unigrams and bigrams and other features. Their performance are listed in Table 14. It shows that the performance of all the algorithms are much lower than the best LR classifier.

Table 14: Performance of Other Final Classifiers on the Test Data.

Model	**F1**	**Recall**	**Precision**
RF	0.648	0.817	0.537
SVM	0.636	0.789	0.533
BT	0.663	0.775	0.579
NN	0.654	0.732	0.591

In summary, we concluded that by following tweets posted in the subsequent 2 years by 'wish to move' users, the LR classifier constructed with unigrams, bigrams and specific other engineered features achieve the highest performance of an F1-score of 0.715. The top 10 most important features

and their coefficients in the best final classifier are listed in Table 15. It can be seen that all the top 10 features are word grams which identify actual relocation. It demonstrates that a user who is going to relocate is more likely to post tweets containing these terms. There are 5 move related terms 'verhuisd (moved)', 'verhuizing (move), 'verhuisdag (moving day)' and 'verhuist naar (moves to)' and 3 first or second person pronouns related terms 'jullie(you)', 'mn (from'm'n', mine)' and 'ik ben (I am)'. Besides the top 10 features, the coefficient of Moved type sentence is 0.816, which is in line with our expectation that relocation users post move related tweets. The feature entity has its coefficient -0.560 in the classifier, which indicates that relocation users likely include more named entities in their tweets. The four features numTwt, txtLen, numW1 and numW1 all show positive coefficients (1.221, 0.551,0.448 and 0.481) in the classifier, showing relocation users incline to post more move related tweets than non relocation ones no matter in terms of the number of tweets or the length of tweets.

Table 15: Top 10 Most Important Features and Their Coefficients in the Best Final Classifier

Feature	**Coefficient**
'verhuisd(moved)'	3.711
'verhuizing(move)'	3.008
'verhuisdag(moving day)'	2.758
'jullie(you)'	2.137
'wordt(is becoming)'	1.587
'mn(from 'm'n'; mine)'	1.494
'ik ben(I am)'	1.409
'nu(now)'	1.405
'zit(is sitting)'	1.332
'verhuist naar(moves to)'	1.328

5 Discussion

5.1 False Positives and False Negatives

False Positives. One of the big challenges in detecting 'wish to move' or actual relocation is to keep the number of false positives as low as possible. The strategy taken includes creating features not only from the point view of the positive class but also from that of the negative class. To be specific, for example, in the 'wish to move' classifier, the engineered feature willMove is to capture 'want to move' type sentences. While the feature nonpersonal is to distinguish institution related move from our target household move. Figure 7 shows daily count of 'wish to move' tweets during the WoON survey identified by the best 'wish to move' classifier. It can be seen that, after duplicates and retweet elimination, the three large peaks are all gone and the daily distribution of 'wish to move' tweets becomes much more stable. It varies between around 30 to 100. In a future application of identifying 'wish to move', this kind of distribution plot, especially the one after elimination can be used to check for the occurrence of any unexpected interfering messages. If there are still abnormal peaks after elimination, it may imply that there are sets of tweets resembling 'wish to move' during a specific period, which could be affected by some trending topic and mistaken by the classifier. In this way, more potential false positives could be discovered and removed. However, there still could be other kinds of unexpected messages included.

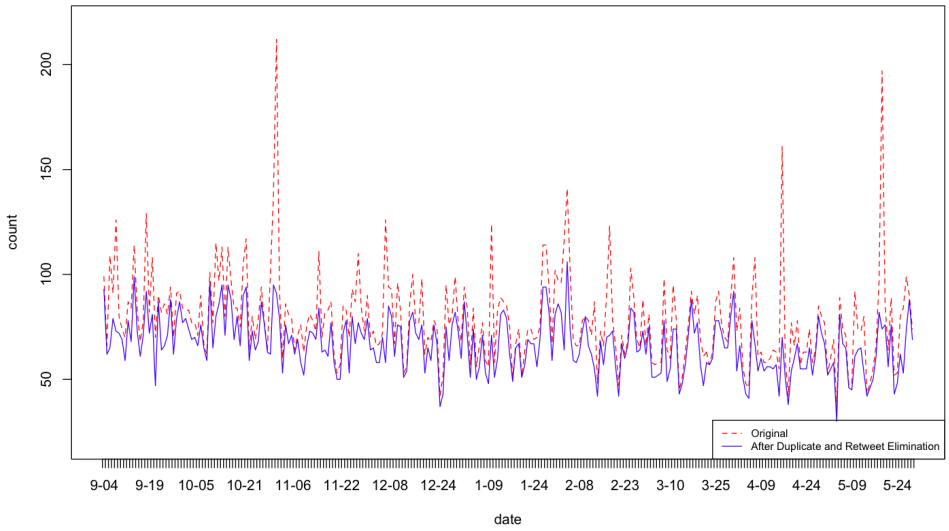


Figure 7: Daily Count of Classified 'Wish to Move' Tweets

In the research, we found that it is quite challenging to identify 'wish to move' or relocation tweets since:

i) We identify them among the keywords based tweet collection, that means, those tweets themselves are alike to each other in terms of those keywords

ii) 'Wish to move' or relocation tweets only account for a small portion of the keyword-based collection (6.91% and 16.2% in our training data, respectively), which directly leads to limited information of the positive class learned by the classifier, even though strategies dealing with imbalanced data were adopted (increasing weight for positive class, stratified training/test split when cross validation and using alternative metrics).

iii) 'Wish to move' and relocation are personal events, unlike tweets of trending topics gathering usually around a specific time point. Tweets indicating personal events naturally spread over the whole time line and do not erupt together. It makes it hard to cluster or identify them by leveraging other features than text, for instance, temporary feature.

iv) Some negative tweets really resemble positive tweets a lot. Just take trigrams 'zijn verhuisd naar (have moved to)' and 'is verhuisd naar (has

moved to)' as examples, they are among top 10 important features in the relocation classifier. Intuitively one might expect they more indicate relocation tweets. However, both of them are found with negative coefficients in the relocation classifier, indicating that there are even more mentions of them in non-relocation tweets. Also some tweets are clearly sarcastic, for example, 'I want to move and preferably to another planet', which does not really indicate a 'wish to move' but resemble them a lot.

Feature Assembling. The classifiers work in a way that they can recognize 'wish to move' or relocation tweets conveyed in frequently used words, word combination or sentence structures. It means that if a 'wish to move' or a relocation tweet is expressed in other uncommon forms, it could be difficult for the classifiers to recognize them. Therefore, it is helpful to engineer features by assembling multiple forms into one feature. Take the feature willMove as an example, it consists of multiple indicative sentence structures e.g. 'move tomorrow', 'want to relocate' and 'want to move', etc. If only taking one single sentence structure as one feature, it will be hard for each structure to become discriminative enough. However, when a feature is created by assembling these multiple sentence structures together, the discriminative power of the assembled feature improves a lot. In this way, even uncommon forms can be caught and become powerful in recognizing tweets so that both false positive and false negative could be reduced. From this perspective, more discriminative features could be engineered in this way in future work to improve the classifiers.

Adding More Positive Cases. Strategies of a stratified training/test data split, increasing weight for minority (positive) class when training and applying alternative metrics were adopted to deal with imbalanced data. Besides, in our research it shows that adding more positive cases does help to improve the performance of a classifier. Another research claims that using balanced training data in binary classification (50% positive and 50% negative) results in the highest accuracy regardless of the proportion of the two classes in test data [34]. Considering the positive class in the training data in our research accounts for far below 50%, adding more positive cases will be an effective action to improve the classifiers developed in the research further.

Query Keyword Selection. When collecting potential 'wish to move' and relocation messages from Coosto, several certain keywords are used to query ('verhuis*' or 'verhuiz*' for 'wish to move' and 'verhuisd' or 'nieuw huis' for

relocation). Those keywords were determined after we experimented with other words and found those brought us target messages with relatively less false positives. It can happen that a user posts his 'wish to move' without using those selected keywords. Therefore, we could lose some target messages due to the limitation of the selected keywords. In future research, other query keywords can be used but at the same time it is important to consider the trade-off between introducing target messages and false positives.

5.2 Population Representativeness

Though 80% of the Dutch population actively participates in social networks, users on those networks differ from the general Dutch population in terms of age and education level [5]. Young people and highly educated people are more active on social networks. Both groups are among the frequent movers in which we are interested. However, the population on social networks is still selective and not everyone will actively post tweets. Besides, the tweets analysed in the research were all publicly available. Twitter users with protected accounts are not included in our research.

There is no similar pattern or positive correlation discovered between the 'wish to move' proportion concluded from the WoON survey and the one derived from Twitter in this research. The analysis of this discrepancy is as follows. First, the currently best 'wish to move' classifier gives a recall of 0.625 and a precision of 0.5, hence there are still 'wish to move' tweets unidentified and non 'wish to move' tweets mistaken as 'wish to move' by the classifier. Second, the 'wish to move' proportion derived from Twitter is calculated as the proportion of 'wish to move' tweets to all the tweets during the same period. The proportion is actually an estimate of the proportion of 'wish to move' users which cannot be derived directly since the number of unique Twitter users during a certain period is unknown. Third, the population behind the WoON survey and on Twitter is likely not same as discussed above. In addition, although on average Twitter users have 41% of their messages about themselves, there are users who do not or seldom share their own personal life on Twitter but just do information sharing, self-promotion, opinions/complaints and so on [26]. For the latter ones, it can be expected that they will not tweet their 'wish to move'.

Per their tweets posted afterwards, more than 20% of the 1,000 'wish to move' users relocated within the subsequent two years. It verifies that many

of the 'wish to move' users are the relocation users. However, it could be some relocation users who just do not post any 'wish to move' ahead. Their relocation can not be identified with our method since our pipeline starts with identifying and tracking 'wish to move' users.

Username Change. Twitter allows its users to change their usernames. Currently in Coosto, a user can only be followed by his username. Hence, if a 'wish to move' user changes the username, the follow up will be lost. That is to say, if a username change happens, then there will be no way for us to know if he will actually relocate. This could introduce bias to the percentage of 'wish to move' users who actually moved if we are interested to calculate that.

Tweet Length Limit. Twitter doubled its character limit from 140 to 280 since November 2017. All the tweets used in this research were posted during 2014 to 2017 and the majority of them were before the extended limit hence within 140 character limit. Although the character limit extended, Twitter mentions that only 5% of tweets exceeded 140 characters. Also, applications of Twitter by users may change over time. Therefore, it would be worthwhile to check, in the future, if the classifiers built in this research are applicable to future tweets, it can be considered to refit them to the newly produced tweets under the extended limit.

Generalization. The pipeline developed in this research has the potential to be generalized to any other personal event detection where discrepancy between people's wish and their actual behaviour is interesting, e.g. marriage, education, graduation and tourism. Once adapting engineered features to specific personal events, the approach developed in this research could be applied to a variety of other personal topics.

6 Conclusions

With certain features (word grams and derived features) of the text of tweets, 'wish to move' or relocation tweets can be distinguished from other tweets. Adding word bigrams and trigrams improves the discriminative power of the classifiers constructed with only unigrams and other engineered features. This is in line with the expectation that unigrams concern only a single word, while bigrams and trigrams provide additional information in form of word combination and sentence structure. However, when taking more than one follow-up tweet to determine if a 'wish to move' user relocates, adding trigrams does not further improve the classifier. A URL more likely appears in non 'wish to move' or non relocation tweets in comparison to 'wish to move' or relocation ones. Other engineered features than word grams like sentiment, text length, person pronoun and specific sentence structure etc. are important in discriminating 'wish to move' and relocation tweets from tweets that do not relate to that. It is found that 'wish to move' or relocation tweets are more likely to express a negative or a positive emotion rather than a neutral one. The longer a tweet, the less likely it expresses 'wish to move' or relocation. Compared with other tweets, there are more first and second person pronouns mentioned in 'wish to move' and relocation tweets. This is consistent with our objective in which we are interested in people who themselves wish to move or actually relocate rather than those who refer to someone else. The sentence structures willMove and notMoved are among top 10 of the most important features respectively in the 'wish to move' classifier and the relocation classifier.

It is concluded that 'wish to move' or relocation messages posted on Twitter contain certain words, word combinations and sentence structures which distinguish them from other tweets. The best 'wish to move' classifiers in the research achieves a recall of 0.625 and a precision of 0.5 and the best relocation achieves 0.778 and 0.438. There is still room for improvement in the classifiers, especially in precision. Defining features by summarizing similarity in false positives could be a good starting point in the further research.

The longer a 'wish to move' user is followed up, the more accurate the classifier can tell if he relocated. By following the tweets posted in two years, relocation can be identified with an F1-score of 0.701. This makes it possible to determine the proportion of 'wish to move' users who actually

relocated in the subsequent period. A similar approach might be applied to improve the classifier for 'wish to move' messages as well.

No similar pattern or positive correlation is found between the weekly 'wish to move' proportion concluded from the WoON survey and the one derived from the Twitter. The discrepancy between the two proportions is discussed above from several aspects. It is concluded that the proportion derived from Twitter in this research could be taken as a very first step to get the 'wish to move' proportion of the Dutch population based on social media data.

References

[1] Antenucci, D., Cafarella, M., Levenstein, M., Ré, C., and Shapiro, M. D. Using social media to measure labor market flows. Tech. rep., National Bureau of Economic Research, 2014.

[2] Becker, H., Naaman, M., and Gravano, L. Beyond trending topics: Real-world event identification on twitter. *Icwsm 11*, 2011 (2011), 438–441.

[3] CBS. Netherlands housing survey (woon).

[4] CBS. Seven in ten internet users active on social media, 2012.

[5] CBS. Ict, knowledge and the economy 2017. Tech. rep., CBS, 2017.

[6] Ceron, A., Curini, L., Iacus, S. M., and Porro, G. Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens political preferences with an application to italy and france. *New Media & Society 16*, 2 (2014), 340–358.

[7] Choudhury, S., and Alani, H. Personal life event detection from social media.

[8] Daas, P. J., Puts, M., Tennekes, M., and Priem, A. Big data as a data source for official statistics: experiences at statistics netherlands.

[9] Daas, P. J., and Puts, M. J. Social media sentiment and consumer confidence. Tech. rep., ECB Statistics Paper, 2014.

[10] Daas, P. J., and Van Beuningen, J. Measuring the wish to move by social media (in dutch). *Internal CBS report* (2018).

[11] Davis, J., and Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning* (2006), ACM, pp. 233–240.

[12] De Groot, C., Mulder, C. H., and Manting, D. Intentions to move and actual moving behaviour in the netherlands. *Housing Studies 26*, 03 (2011), 307–328.

[13] Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*, vol. 1. Springer series in statistics New York, NY, USA:, 2001.

[14] Gomide, J., Veloso, A., Meira Jr, W., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the 3rd international web science conference* (2011), ACM, p. 3.

[15] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., and Donaldson, L. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. *BMJ Qual Saf 22*, 3 (2013), 251–255.

[16] Han, J. *Data mining : concepts and techniques*. Elsevier Morgan Kaufmann, Amsterdam Boston San Francisco, CA, 2006.

[17] Hürriyetolu, A., Gudehus, C., Oostdijk, N., and van den Bosch, A. Relevancer: Finding and labeling relevant information in tweet collections. In *International Conference on Social Informatics* (2016), Springer, pp. 210–224.

[18] Ionescu, R. *Knowledge transfer between computer vision and text mining : similarity-based learning approaches*. Springer, Switzerland, 2016.

[19] Kunneman, F. *Modelling patterns of time and emotion in Twitter# anticipointment*. PhD thesis, Sl: sn, 2017.

[20] Kunneman, F., and van den Bosch, A. Event detection in twitter: A machine-learning approach based on term pivoting.

[21] Kunneman, F., and Van den Bosch, A. Open-domain extraction of future events from twitter. *Natural Language Engineering 22*, 5 (2016), 655–686.

[22] Lassen, N. B., la Cour, L., and Vatrapu, R. Predictive analytics with social media data. *The SAGE Handbook of Social Media Research Methods* (2017), 328.

[23] Lobo, J. M., Jiménez-Valverde, A., and Real, R. Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography 17*, 2 (2008), 145–151.

[24] McHugh, M. L. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica 22*, 3 (2012), 276–282.

[25] McMinn, A. J., Moshfeghi, Y., and Jose, J. M. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (2013), ACM, pp. 409–418.

[26] Naaman, M., Boase, J., and Lai, C.-H. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (2010), ACM, pp. 189–192.

[27] Nguyen, D., Trieschnigg, D., and Meder, T. Tweetgenie: Development, evaluation, and lessons learned. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations* (2014), pp. 62–66.

[28] Qualman, E. *Socialnomics: How social media transforms the way we live and do business*. John Wiley & Sons, 2010.

[29] Rice, D. M. Is the auc the best measure?

[30] Ripberger, J. T., Jenkins-Smith, H. C., Silva, C. L., Carlson, D. E., and Henderson, M. Social media and severe weather: Do tweets provide a valid indicator of public attention to severe weather risk communication? *Weather, Climate, and Society 6*, 4 (2014), 520–530.

[31] Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., and Conrad, F. G. Social media analyses for social measurement. *Public opinion quarterly 80*, 1 (2016), 180–211.

[32] van den Brakel, J., Söhler, E., Daas, P., and Buelens, B. Social media as a data source for official statistics; the dutch consumer confidence index. *Survey Methodology 43*, 2 (2017).

[33] Walther, M., and Kaisser, M. Geo-spatial event detection in the twitter stream. In *European conference on information retrieval* (2013), Springer, pp. 356–367.

[34] Wei, Q., and Dunbrack Jr, R. L. The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PloS one 8*, 7 (2013), e67863.